DOCUMENT RESUME

ED 156 719 95 TM 007 286

AUTHOR House, Ernest R.
TITLE The Logic of Evaluative Argument. CSE Monograph Series in Evaluation, 7.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 77
CONTRACT 400-77-0034
NOTE 69p.
AVAILABLE FROM Center for the Study of Evaluation, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California 90024 ($4.50)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Abstract Reasoning; Audiences; Bias; Case Studies; Credibility; Data Analysis; Decision Making; *Evaluation; Evaluation Methods; *Evaluation Needs; Evaluative Thinking; Evaluators; Logic; *Logical Thinking; Mathematical Models; *Models; *Persuasive Discourse; Problem Solving; Responsibility; Summative Evaluation; *Values
IDENTIFIERS Glass (Gene V); Scriven (Michael)

ABSTRACT

Evaluation is an act of persuasion directed to a specific audience concerning the solution of a problem. The process of evaluation is prescribed by the nature of knowledge--which is generally complex, always uncertain (in varying degrees), and not always propositional--and by the nature of logic, which is always selective. In the process of persuasion one must ascertain who the audience is and find a basis of agreement on premises, both of facts and values, and on presumptions. Two criteria for evaluation are: the most efficient way to a given end, or the most effective use of available resources. Quantitative evaluation methods involve three stages: (1) substantive definition of the problem and its translation into a formal, mathematical model; (2) compilation of information in terms of the formal model and its formal, logical analysis; and (3) translation of the formal conclusions back into substantive terms. Both formulation and interpretation require good intuitive judgment. The evaluator and the audience must employ their reasoning in a dialogue, and both must assume responsibility, since evaluation is never completely convincing nor entirely arbitrary. The logical arguments used in two works are discussed. The works--Gene V. Glass' review of Michael Scriven's instructional cassette lecture on "Evaluation Skills;" and Scriven's reply--are appended. (Author/CTM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mark Young

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED156719

TM007 286

CSE MONOGRAPH SERIES IN EVALUATION

7

# THE LOGIC OF EVALUATIVE ARGUMENT

Ernest R. House

CENTER FOR THE STUDY OF EVALUATION
UNIVERSITY OF CALIFORNIA · LOS ANGELES

# CSE MONOGRAPH SERIES IN EVALUATION

SERIES EDITOR

Eva L. Baker

Center for the Study of Evaluation
UCLA Graduate School of Education
University of California, Los Angeles
Los Angeles, California 90024

## CSE MONOGRAPH SERIES IN EVALUATION

NUMBER

1. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case from the MINNEMAST Project
   Wells Hively, Graham Maxwell, George Rabehl, Donald Sension, and Stephen Lundin $3.50

2. National Priorities for Elementary Education
   Ralph Hoepfner, Paul A. Bradley, and William J. Doherty $3.50

3. Problems in Criterion-Referenced Measurement
   Chester W. Harris, Marvin C. Alkin, and W. James Popham (Editors) $3.50

4. Evaluation and Decision Making: The Title VII Experience
   Marvin C. Alkin, Jacqueline Kosecoff, Carol Fitz-Gibbon, and Richard Seligman $3.50

5. Evaluation Study of the California Preschool Program
   Ralph Hoepfner and Arlene Fink $3.50

6. Achievement Test Items—Methods of Study
   Chester W. Harris, Andrea Pearlman, and Rand R. Wilcox $4.50

7. The Logic of Evaluative Argument
   Ernest R. House $4.50

This project has been funded at least in part with Federal funds from the Department of Health, Education, and Welfare under contract number NIE 400-77-0034. The contents of this publication do not necessarily reflect the views or policies of the Department of Health, Education, and Welfare, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

## TABLE OF CONTENTS

Foreword . . . . . vii

Chapter I: Evaluation as Argument . . . . . 1

The Coming Great California Earthquake
Equivocality of Evidence: Certainty vs. Credibility
Evaluation as Persuasion
The Evaluation Audiences
Premises of Agreement
Quantitative Argument
Qualitative Argument
Ambiguity and the Development of an Argument

Chapter II: The Logic of the Argument . . . . . 22

Modes of Reasoning
Analysis of Glass's "Educational Product Evaluation"
Analysis of Scriven's Response to Glass's Evaluation
Naturalistic Evaluation
Objectivity, Validity, and Impartiality Reconsidered
Evaluative Discourse: The Good Life (along the San Andreas Fault)

References . . . . . 48

Appendix

Glass's Evaluation—Educational Product Evaluation: A Prototype Format Applied . . . . . 53

Scriven's Response to Glass—Educational Product Re-Evaluation . . . . . 61

# FOREWORD

The institution of schooling, existing as it does within a society undergoing rapid changes, has multiple problems and only limited resources of money, facilities, manpower, skills, and information to meet them. Good information to guide both policy making and management is an essential resource at all levels of educational decision making but is especially important at the local district level; for it is here that choices about curriculum, instruction, and delivery of other educational services most directly affect the daily actions of principals, teachers, and students.

There are those who assert that evaluation activities, if they are based on a broad set of methodologies derived from the social sciences, can provide valid, reliable, and relevant information for a range of educational decisions—instructional, curricular, managerial, policy making. Up until now, admittedly, evaluators have not been able to provide this information in a relevant and timely manner. In the ten years or so since evaluation has been formally called upon to bear burdens both for accountability and policy making, the limitations of the technology available to evaluators have become painfully apparent. If evaluation can be considered a discipline, it is one which grows by accretion; the agenda of unsolved problems, both theoretical and practical, attracts researchers from diverse disciplines who apply diverse methodologies and call their work evaluation.

The mission of the Center for the Study of Evaluation (CSE) is to study, from a variety of perspectives, the act of evaluation as it affects educational programs and services.

The case for evaluation is based on the premise that people will, if they can, make changes based upon information. It is thus assumed that decisions about policies, programs, students, or services will be more rational if good information is available when needed. The simplicity of this idea has been challenged by many who view decision making as the result of influences far more diffuse than those within the conscious control of the decision maker. They hold that political, social, psychological, or organizational factors, while often unarticulated, dominate the decision process. This monograph explores the act of decision making from an analytical perspective.

Dr. House was a resident Visiting Scholar at CSE in 1976. During that period, he worked alongside staff, provided counsel on a variety of problems, and prepared the monograph presented here. Participants in the Visiting Scholar Program include recognized scholars in the conceptual and policy making areas of evaluation as well as methodologists primarily concerned with the design, analysis, and interpretation of empirical studies. Members of the practitioner community are also invited to share

their perceptions of how CSE activities might assist school people in their evaluation tasks.

This monograph approaches the analysis of evaluation from perspectives that have little reliance on quantitative origins. We welcome Dr. House's point of view and expect it to generate discussion within the field. It is our intent, through publications such as this, to stimulate the membership of the field of evaluation to expand or to consolidate positions related to the purposes, methods, and uses of educational evaluation. We look forward to your comments.

Eva L. Baker
Director
CSE

## Chapter I

# EVALUATION AS ARGUMENT

> I choose the word "argument" thoughtfully, for scientific demonstrations, even mathematical proofs, are fundamentally acts of persuasion. Scientific statements can never be certain; they can only be more or less credible.
>
> Joseph Weizenbaum
> in *Computer Power and Human Reason*, 1976.

> Generalizations decay.
>
> Lee J. Cronbach
> in *Beyond the Two Disciplines of Scientific Psychology*, 1974.

## THE COMING GREAT CALIFORNIA EARTHQUAKE

I sit in Los Angeles but wonder why I stay. A sudden one-foot uplift has appeared along a hundred-mile strip of the San Andreas fault. Based on seismic wave readings, a California scientist has predicted a major earthquake for the Los Angeles area within a year (*Science*, May 1976). Based on different readings, a radio evangelist warns of a major quake. Both scientists and seers agree in their prophecies. Neither provides the kind of information I need.

I talk to the natives about these ominous signs. Their response is shaped by the necessity of living in such circumstances; they shrug their shoulders. The President has been informed, but no one seems to know exactly what to do. Washington officials suggest setting up a new array of scientific instruments along the fault, although what will result from more measurement is not clear.

Meanwhile the weather is perfect, the setting in the Santa Monica Mountains splendid, the lifestyle sybaritic. Calculations of probabilities of long-term seismic events do me no good; I need to know when the earth will move in relation to myself.

The vocabulary of action is complex. Everyone agrees that information somehow informs decisions but the relationship is not direct, not simple. Often the more important the decision, the more obscure the relationship seems to be. Consider the decision to marry. For most people, it is a long, arduous process, one which takes shape over a period of time. No single piece of information serves as a decision-point. Quite the contrary. The decision proceeds slowly, almost imperceptibly, until it arrives. Reason after reason is advanced and tried out. Finally, a multiplicity of arguments serves as a rationale for the decision, which is often made long before all the arguments are advanced.

I wish to thank Lee Cronbach, Bob Ennis, Gene Glass, and the CSE staff for detailed comments on the monograph.

The most significant decisions are those that have long-range implications but defy easy extrapolation, that are so entangled with everything else that they resist precise formal analysis. To those we are forced to apply our intuitive logic, our common sense. It is in the nature of these complex problems that knowledge about them is limited, that it is less than determinate. In the face of uncertain knowledge, the task of entangled decision making becomes less one of absolutely convincing ourselves with proofs than one of persuading ourselves with multiple reasons. The criterion becomes not what is necessary but what is plausible.

## EQUIVOCALITY OF EVIDENCE: CERTAINTY VS. CREDIBILITY[1]

Why, then, do government officials, the public, and even members of the evaluation community call for definitive proof of the success of educational programs? There is a tradition as old as Descartes which says that the *only* knowledge is that which is certain. Descartes's method of analysis was one of total skepticism: to doubt everything that could be doubted. In his search for certain knowledge, he arrived at the *self-evident* as the ultimate mark of reason. For something to qualify as knowledge it had to start from clear and distinct ideas and be extended by deductive proofs. Propositions so derived were thus necessary and compelling to the intellect; they could not be rationally denied.

This method excluded the merely credible from consideration as knowledge. In the Cartesian ideal, the only true reasoning is analytic. Formal deductive logic, the method of proof used in mathematics, is the method par excellence. Knowledge can be reduced to self-evident propositions. In certain knowledge there can be no disagreement. As Descartes wrote, if there is disagreement over a matter between two men, one of them must surely be wrong. There is a true and a false, and logic works by compelling proofs to determine which is which.

Later, those who pursued this line of reasoning confronted the fact that rational men often seemed to reason differently and arrive at contradictory conclusions. Some of Descartes's own propositions looked particularly suspicious. Pascal introduced the explanation that such disagreement as well as the reluctance to accept necessary conclusions was a result of irrationality. Man was seen to possess an irrational side which often led him astray in his search for knowledge. The apparent irrationality of those who do not accept conclusions which others perceive as compelling is a common motif in contemporary evaluation.

From the Cartesian perspective, certain knowledge can be obtained primarily by deductive processes and it must lead to absolute conviction.

[1]For this distinction and many other ideas in this paper, I am indebted to Perelman and Olbrechts-Tyteca's excellent modern work on argumentation *The New Rhetoric: A Treatise on Argument*. University of Notre Dame Press, 1969. 566 pages.

Such reasoning may work in geometry, but it does so by excluding most of the sensate world. As Hume pointed out, our beliefs, even in concepts as basic as causality, are not certain when a thorough skepticism is applied to them. Deductive reasoning succeeds in producing certain knowledge primarily by eliminating most of the everyday world.

The sensate world was epistemologically salvaged for our use by John Stuart Mill. Just as logicians had constructed formal deductive logic by reflecting on the nature of mathematical proofs, Mill reflected on the associationist psychology of his time and formulated an inductive logic that purported to introduce certainty into inductively-derived knowledge. To do this Mill made several assumptions that still pervade survey research today. According to Hamilton (1976), the axioms include the following:

- There is a uniformity of nature in time and space. This lends to inductive reasoning the same procedural certainty as to conclusions drawn from syllogistic logic.
- Concepts can be defined by direct reference to empirical categories and laws of nature can be inductively derived from data because of the above.
- Large samples can suppress idiosyncracies and reveal "general causes."
- The social and natural sciences have the same aim of discovering general laws (which provide a basis for explanation and predictions).
- The social and natural sciences are methodologically identical.
- The social sciences are merely more complex.

Thus, Mill contended that certain knowledge was derivable from inductive reasoning as well as from the deductive. One could define categories and relate them to each other by now familiar techniques. In fact, Mill concluded that the inductive method was the *only* way of discovering new ideas since deductive logic could only reveal what was already there. (Mill was so certain of his method that he contended that ethical principles could also be derived by inductive reasoning and hence had a scientific base.)[2]

[2] I have discussed the powerful effect utilitarian ethics has had on the practice of evaluation in a paper entitled, "Justice in Evaluation" in *Evaluation Studies Annual Review*, Gene Glass, editor, Beverly Hills, CA: Sage Publications, 1976. At the end of his masterpiece on inductive logic, Mill considers the logic of a "practice" or "art." "There must be some standard by which to determine the goodness or badness, absolute and comparative, of ends, or objects of desire. And whatever that standard is, there can be but one; for if there were several ultimate principles of conduct, the same conduct might be approved by one of those principles and condemned by another; and there would be needed some general principle, as umpire between them." John Stuart Mill in *A System of Logic*, Harper, New York, 1893 (8th Edition).

This leads Mill to impose a single universal standard by which to judge practical affairs, for the only alternative is by "supposing a moral sense or instinct" or "intuitive moral principles." General ethical principles can only be known by induction. Since inductive

Mill's first assumption is the important one. In Mill's own words, "The universe, so far as known to us, is so constituted, that whatever is true in one case, is true in all cases of a certain description; the only difficulty is, to find what description" (Mill, 1848). How familiar that idea is to anyone who has engaged in survey research and how fallible the inductive logic on which it is based!

The procedure of reasoning from "some" to "all" is clearly a logical fallacy. Each confirming instance is supposed to make a hypothesis more likely. Yet if the hypothesis is "All men are less than 100 feet tall" and one finds a man 99 feet tall, this is a confirming instance that weakens the hypothesis considerably rather than strengthens it (*Scientific American*, March, 1976). Does every day that goes by in Los Angeles without the predicted great quake make it more or less likely? It is also quite possible in statistical studies to confirm a hypothesis by two independent studies and yet disconfirm the hypothesis by using the total results of the two studies taken together. (See Simpson's paradox in Martin Gardner, "Mathematical Games," *Scientific American*, 1976.)

Nonetheless, in spite of serious flaws of logic, "science" based on inductive logic seems to work with some degree of success. Certainty of knowing, however, is lacking. Even the best established scientific facts must be held as tentative. As one scientist put it:

> The man in the street surely believes such scientific facts to be as well-established, as well-proven, as his own existence. His certitude is an illusion. Nor is the scientist himself immune to the same illusion. In his praxis, he must, after all, suspend disbelief in order to do or think anything at all. He is rather like a theatergoer, who, in order to participate in and understand what is happening on the stage, must for a time pretend to himself that he is witnessing real events. The scientist must believe his working hypothesis, together with its vast underlying structure of theories and assumptions, even if only for the sake of the argument. Often the "argument" extends over his entire lifetime. Gradually he becomes what he at first merely pretended to be: a true believer. I choose the word "argument" thoughtfully, for scientific demonstrations, even mathematical proofs, are fundamentally acts of persuasion.
>
> Scientific statements can never be certain; they can be only more or less credible. And credibility is a term in individual psychology, i.e., a term that has meaning only with respect to an individual observer. To say that some proposition is credible is, after all, to say that it is believed by an agent who is free not to believe it, that is, by an observer who, after exercising judgment and (possibly) intuition, chooses to accept the proposition as worthy of his believing it (Weizenbaum, 1976).

---

certainty pre-supposes a uniformity of nature, the resultant psychology is deterministic. Morality is natural since only a naturalistic assessment will allow scientific methods of proof. Hedonistic utilitarianism is the only basis.

In a sense, Mill was preventing disagreement over moral issues since it is always possible to reach opposite conclusions when there is no previous agreement on a criterion. The result of this reasoning is utilitarian calculation which conflates all human desires into a single configuration and satisfies them by the criterion of maximum total satisfactions derived. The judging is done by an "impartial spectator," who in modern times demonstrates his impartiality by employing "objective" techniques of analysis.

## EVALUATION AS PERSUASION

If even demonstrations in the physical sciences are fundamentally acts of persuasion, inquiries in education are more so. Mill's assumption that the social and natural sciences are methodologically identical seems much more dubious today. Cronbach (1974), for one, doubts the advisability of imposing physical science ideals in social science. In the physical science paradigm, events are explained and predicted by "a network of propositions connecting abstract constructs."

After reviewing twenty years of aptitude treatment interaction studies, which were based on such a model, Cronbach concluded that social phenomena are too open to interactions with other variables to support stable generalizations. The positivistic strategy of fixing conditions in which to reach generalizations assumes steady processes that can be separated into independent systems for study, a fragile assumption in social systems.

Cronbach has suggested interpreting data in context rather than trying to arrive at generalizations. An observer in a particular setting can describe and interpret effects within local conditions. Whereas experimental control and systematic correlation ask formal questions in advance, local observation is more open to the unanticipated. Short-term empiricism is sensitive to the context. In being context sensitive, the researcher may give up some predictive power. He gives up constructing generalizations and theory-building and instead develops "concepts that will help people use their heads." So Cronbach contends.

Evaluations themselves, I would contend, can be no more than acts of persuasion. Although sometimes evaluators promise Cartesian proof and use J. S. Mill's methods of induction, evaluations inevitably lack the certainty of proof and conclusiveness that the public often expects. The definitive evaluation is rare, if it exists at all. Even a scientific methodologist as sophisticated as James Coleman is faced with continued and trenchant criticism of his work. Subjected to serious scrutiny, evaluations always appear equivocal.

Expecting evaluation to provide compelling and necessary conclusions hopes for more than evaluation can deliver. Especially in a pluralistic society, evaluation cannot produce necessary propositions. But if it cannot produce the necessary, it can provide the credible, the plausible, and the probable. Its results are less than certain but still may be useful.

Proving something implies satisfying beyond doubt the understanding of a universal audience with regard to the truth. To produce proof that a universal audience comprised of all rational men would accept requires overcoming local or historical particularities. Certainty requires isolating data from its total context as, for example, in the terms of a syllogism. Logical certainty is achievable only within a closed, totally defined system like a game.

If evaluation is limited to certain knowledge provided by strict deductive

and inductive reasoning, it must abandon a great amount of reasoning power that people ordinarily use in the conduct of their lives. Such a limitation results from confusing rationality with logic. They are not identical.

If absolutely convincing all rational men is too heavy a burden for evaluation, persuading particular men is not. In place of the compelling propositions derived from rigorous logic, one may substitute the non-compelling arguments of persuasion. In place of the necessity of self-evidence, one may substitute variable adherence to theses as presented to particular audiences. The thesis may be more or less credible. The audience is free to believe or not believe after inspecting the arguments and exercising its own judgment.

Persuasion aims at winning a particular audience to a point of view or course of action by an appeal to the audience's reason and understanding. For this purpose, uncertain knowledge is useful although the ideas themselves are always arguable. The appropriate methods are those of argumentation, which is the realm of the "credible, the plausible, and the probable" rather than the necessary (Perelman & Olbrechts-Tyteca, 1969).

Argumentation is contrasted to demonstration. Demonstrations rest on formal logic which avoids ambiguity by the internal consistency of its symbol system. In deductive logic the origin of the axioms is extraneous. When one moves from deduction to induction, all manner of issues become arguable, such as the validity of measurement. But the search is still for "certain" knowledge.

In evaluation, the social and psychological contexts become particularly relevant and the knowledge less certain. Under those conditions argumentation aimed at gaining the adherence and at increasing the understanding of particular audiences, is more appropriate. Persuasion claims validity for only particular audiences and the intensity with which particular audiences accept the evaluative findings is a measure of this effectiveness. The evaluator does not aim at convincing a universal audience of all rational men with the necessity of his conclusions.

Persuasion is directly related to action. Even though evaluation information is less certain than scientific information addressed to a universal audience, persuasion is effective in promoting action because it focuses on a particular audience and musters information with which this audience is concerned. Personalized knowledge that induces people to stop smoking may be different from scientific generalizations linking smoking to heart disease or cancer. Finding out about the heart attack of a close relative is more likely to induce one to exercise than are charts and tables. Evaluative argument is at once less certain, more particularized, more personalized, and more conducive to action than is research information.

In summary, evaluation persuades rather than convinces, argues rather than demonstrates, is credible rather than certain, is variably accepted rather than compelling. This does not mean that it is mere oratory or

entirely arbitrary. Because it is not limited to deductive and inductive logics does not mean that it is irrational. Rationality is not equivalent to logic. Evaluation employs other modes of reasoning. Once the burden of certainty is lifted, the possibilities for informed action are increased rather than decreased.

**CHART 1:**

**Contrasts Between Evaluation as Argumentation and Evaluation as Demonstration**

| *Evaluation as Argumentation* | *Evaluation as Demonstration* |
|---|---|
| Persuasion | Absolute conviction |
| Credibility | Certainty |
| Non-compelling | Necessary |
| Variable adherence | True or false |
| Particular audience | Universal audience |
| Dialectical reasoning | Analytic reasoning |
| Informal logic | Formal logic |
| Reflective | Calculative |
| Action-oriented | Theory-building |
| Tacit knowledge | Explicit knowledge |
| Knowledge in heads | Knowledge in propositions |
| Ambiguous | Clear and distinct |
| Concrete | Abstract |
| Arguable | Definitive |
| Direct experience | Indirect indicators |

## THE EVALUATION AUDIENCES

If persuasion becomes the aim of evaluation, the audiences to whom the evaluation is addressed are important. For years evaluators have been counseled to think of their audiences and the kind of information the audiences will need. What is relevant for one group may not be relevant for another. Argumentation presupposes that a "community of minds" exists, that there is intellectual contact, and that there is agreement on at least a few issues on which deliberation is to begin.

There must be a common language and a desire on the part of the evaluator to persuade the audiences and to take their concerns seriously. Often these conditions are not met. The audiences are misconceived or not

taken seriously. It is not uncommon for the evaluator to muster information appropriate to an audience of psychologists but which has little meaning to a teacher or a government official.

There are at least three general types of audience: the universal audience, a single audience with whom one engages in dialogue, and oneself as an audience. Argumentation with a universal audience strives to gain the adherence of every rational person. Conceptually the universal audience consists of all men at all times so the arguments must be timeless and free of context.

The agreement of a universal audience is likely to be secured by formal logical reasoning based on self-evident concepts. Thus the tighter the experimental design, the more convinced a far-removed universal audience will be of the cause and effect relationship, regardless of the context. A particular audience closer to the scene may assume cause and effect without such proof. Of course, the universal audience is not "aggregatable" at any given time but various elite groups in fact serve as a surrogate for it. Perhaps philosophers more than most represent this type of audience. The arguments that move philosophers are not always the same as those that move teachers.

The more an argument is directed toward a universal audience, the less "arguable" it is. There is little to argue about in pure deductive logic. Evaluation techniques are often presented as being non-argumentative, as, for example, being based on valid and reliable instruments, as employing sound statistical procedures, and so on. In fact, all statements made on the basis of an evaluation are subject to challenge and are arguable—if properly challenged. The more technical and quantitative the evaluation, the less a naive audience will be able to challenge it and the evaluation will appear to be more certain than it is.

In evaluations using statistical metaphors, one can argue that treatment effects differ because there is a probability that two mean test scores belong to different populations and, hence, that the experimental program is better than the control. The extensive use of numbers in the statistical procedures and the test scores gives a semblance of certainty and unequivocality to evidence.

Actually many assumptions lie concealed behind the numbers (as indeed behind every evaluation). One can almost always challenge the validity of the tests, the appropriateness of the statistical procedures, and the control of the experimental design. The challenge does not invalidate the evaluation. But once the premises are challenged, the nature of the evaluation as argumentation becomes apparent. The evaluator may defend his study either successfully or unsuccessfully. In any case, he must resort to non-deductive and more equivocal reasoning if he is to defend it. Although the evaluation has the appearance of appealing to the definitive rationality of the universal audience, it ends in direct appeals to particular audiences. I believe it is impossible to construct an evaluation otherwise.

Even a broad-based evaluation operation like *Consumers Report*, which uses "objective" procedures and sophisticated experimental designs to evaluate consumer products, is an appeal to particular audiences. Its arguments, directed at the upper-middle class, have little meaning for either the lower classes or the upper classes, and its evaluations are little heeded by them.

Thus the situation the evaluator faces is almost always an appeal to particular audiences which he can define with some precision. If he cannot define his audiences, the evaluation is indeterminate. He must address issues and construct arguments that appeal to particular audiences. Furthermore, the audiences are likely to be a composite of several groups which complicates his task considerably. Effective appeal to particular audiences changes the limits of applicable rationality. One is not confined to the most restrictive modes of reasoning. If evaluation becomes more equivocal, it also becomes more possible.

One ideal of two-party argumentation is embodied in the Socratic dialogue. The dialogue develops as a rigorous chain of reasoning between a questioner and a responder. The one-person audience is persuaded by getting him to agree on certain principles point by point. The audience's particular concerns are ultimately addressed in the interaction. The Socratic dialogue is also powerful to third parties who might read it (see Scriven's goal-free dialogue, Scriven, 1973).

The actual audience most evaluators face seldom consists of one person, however. It is most often a composite. Some evaluation theorists have suggested modes of evaluation in which the evaluator engages in frequent exchange with the audience throughout the study (see Stake's 1973 "responsive evaluation" in which the evaluator is expected to respond to the concerns of the program personnel). Whatever the mode of evaluation, I would contend that evaluation which succeeds in being persuasive must engage the audience in fundamental discourse, although that discourse may occur in different ways.

Discourse conducted in this fashion is more than a mere debate in which different points of view are presented by partisans. The dialogue must be a discussion in which the parties seriously and honestly search for mutual answers. This restriction severely qualifies the use of adversary methods as persuasive devices since one may adjudicate a conflict without persuading anyone of anything.

Legal procedures are important new means of encouraging evaluative discourse (Wolf, 1974). Yet to be a successful discourse in which people listen to one another, as opposed to a forensic contest, the acrimony in court trials must be reduced.[3] One must avoid the bias sometimes evident

[3]Ramsey Clark, the former U. S. Attorney General and a trial lawyer, is opposed to the adversary process as a truth-discovering mechanism. "If there is a worse procedure for discovering the truth, I don't know what it is." He claims that no one knows any more after a criminal trial than before. The trial is simply a dramatization for the benefit of the jury. Wolf and Farr (1976), as adversaries in their evaluation of the Indiana University alternative

in courts of law. Admittedly, the distinction between a discourse for discovering truth and mere oratory is not easy to make.

There is at least one other audience one can address in argumentation —oneself. Some have reasoned that arguments addressed to oneself are more likely to be valid and sincere since there is little advantage to fooling oneself. If the "self" is conceived as the program staff, this means formative evaluation. I have seen few really successful formative evaluations. Either the information the evaluator collects is irrelevant to the program staff or the evaluator is perceived as being too much of an outsider to be a credible source.

Kemmis (1976) recently advocated "evaluation as self-criticism." He sees the primary audience as being the program staff itself. Believing a dialectic between knowledge and action to be the only way to improve practice, he suggested that evaluation standards be derived from the program participants themselves and that the data consist of the progress as seen by participants. Evaluation thus becomes therapeutic self-criticism. The ultimate goal is increased understanding and insight of the participants themselves, which can then lead to effective action.

A follower of J. S. Mill would not think highly of this approach since self-knowledge in his viewpoint would be likely to lead to rationalization rather than to reason. Mill thought in terms of the self as audience only insofar as it represented the universal audience. Propositions would be established as either true or false. A more argumentative approach aims at increasing the adherence of the audience rather than demonstrating truth or falsity.

In fact, the difference in viewpoints is more fundamental. It is partially a difference as to where knowledge exists. Does it exist in propositions whose truth can be certified, or does it exist only in individual heads? The view taken here is that knowledge exists only within the mind. The goal of evaluation is not to arrive at a formal statement except as it stimulates understanding in the mind of the audience.

In the argumentative approach, the audience must also share responsibility. Since the information is not compelling, the audience is free to choose its own degree of commitment. It must actively choose how much it wishes to believe. This requires an active testing of the evaluation by the audience itself rather than a passive acceptance or rejection. The audience must make a personal commitment and share responsibility. This rational decision belongs to the audience, not to the evaluator.

## PREMISES OF AGREEMENT

The development of an evaluation argument presupposes agreement on the part of the audiences. The premises of the argument are the beginning of this agreement and the point from which larger agreement is built. Just

---

teacher education program, were well aware of this difficulty and tried to reduce the competition accordingly.

as common sense admits unquestioned truths that are beyond discussion, some of the major premises of an evaluation are tacit rather than explicit.

According to Perelman and Olbrechts-Tyteca (1969), there are two classes of premises: the "real" and the "preferable." The real includes facts, truths, and presumptions and generally claims validity vis-a-vis the universal audience. On the other hand, the preferable is identified with a particular audience and includes values, composite value hierarchies, and value premises of a very general nature called "loci."

Facts and truths are those data and notions which are seen as agreed upon by the universal audience, i.e., held in common by thinking beings, and hence needing no justification. Whether a datum is a fact depends upon one's conception of the universal audience. If the audience changes, so can facts and truths. However to hold the status of a fact or a truth means that for the purposes of argument the datum is noncontroversial and uncontested. If the datum is questioned, it loses its status as a fact and becomes itself an object of argument rather than an object of agreement.

Where there is agreement on the conditions for verification as in modern science, there can be many facts. Many data are not accorded the status of "facts" by modern science. Polanyi (1958) pointed out how science protects its own system of beliefs from inconsistency by denying various data as factual which conflict with other beliefs. Thus for many years science did not recognize hypnotic effects as occurring at all. These data were not recognized as factual because they conflicted with the current general scientific belief system. This belief system may change from time to time, but regardless of what it excludes, arguments within the belief system must be based on uncontested facts and truths.

Arguments also proceed from presumptions which do not have the full authority and confidence of a fact or truth. Presumptions cannot be proved but are nonetheless widely accepted as being tentatively true. Many presumptions are connected to the concept of the normal. In evaluations employing statistical models and metaphors, the assumption that attributes within a population are normally distributed is almost universally accepted. Perhaps an implicit presumption of all evaluations is that the act of evaluation itself will somehow improve the program under inspection.

The second class of objects of agreement is that of the preferable. Objects of preference claim the adherence of only particular groups rather than that of the universal audience. Values are the most conspicuous examples. Agreement with regard to a value is an admission that there is a specific influence on action or a disposition toward action that the evaluator can make use of. Although relevant for a particular group, a value is not regarded as binding on everyone.

In science, values enter primarily in the selection of objects of interest for investigation since one cannot investigate the entire world (Polanyi,

1958) and possibly in the acceptance of scientific conclusions by overall human judgment (Weizenbaum, 1976). But during most of the argument, especially in the exact sciences, values are supposed to be excluded. Ennis's (1973) analysis of cause and effect relationships leads one to question this. In evaluation there is no question that values enter at every stage. Values are used to persuade the audiences and to justify choices to others.

Abstract values like truth, beauty, and justice have a universal appeal only because they are so general and unspecified. Once their content is determined, they appeal to certain audiences and not to others. Their role is to justify choices where there is *not* unanimous agreement. For example, in my analysis of justice in evaluation (House, 1976), I contrasted three specific conceptions of justice: the utilitarian, the pluralist-intuitive, and justice-as-fairness. The purpose of the analysis was to justify protecting people being evaluated and to promote more egalitarian criteria in actual evaluations. The analysis was warmly endorsed by those who agreed with such values and was not well accepted by those who did not, although everyone is in favor of "justice."

Abstract values like justice can be contrasted with concrete values like America or individual persons. Abstract values are more readily used for criticisms as they are not respectors of individual persons. Concrete values like fidelity and solidarity lend themselves more to compromise and conservative argument. Of course values are not held exclusively by any group. Audiences are perhaps better characterized by the relative weights given to various values.

Various combinations of arguments can be compressed into a few general groupings called "loci" (Perelman & Olbrechts-Tyteca, 1969). The most common loci are those of quantity and quality. Arguments grouped around the loci of quantity affirm that one thing is better than another for quantitative reasons: greater number, higher degree, more durability, etc. The effectiveness of means will often be justified by quantitative loci. The idea of the normal and the norm are also based on quantity.

Contrasted with quantity is the idea of quality. Something has high value even though it defies number. Associated with quality is a high rating of the unique. One can be in possession of truth while the multitude is in error. For example, Scriven (1972) contended that the notion of objectivity is not necessarily linked to the number of people holding an idea, nor subjectivity to one person's perception, as is often believed.

Besides general agreements on facts and values, there are special agreements particular to certain special audiences and particular to each evaluation. To the extent that the evaluation is addressed to a technical audience, that audience will share certain agreements and conventions. A group of educational researchers is such a technical audience. Evaluations directed toward a lay audience cannot rely on the same agreements.

Perhaps the most important agreements peculiar to a particular evaluation are those derived from the negotiation that often precedes the evaluation—agreements between sponsors, program personnel, and evaluators. In this exceedingly important negotiation, agreement can be reached on criteria, methods and procedures, access, dissemination of results, and so on. Disagreement on these points can destroy the entire credibility of the evaluation.

In summary, at the beginning of an evaluation, the evaluator must build upon agreements with the audiences. These agreements may be implicit as well as explicit. In fact, it would be impossible to specify all these understandings, although it is quite dangerous to assume agreement on important points where there is none. The evaluator must start from where his audiences are, even though the beginning premises may not be acceptable to other parties nor to the evaluator himself. Otherwise the evaluation will not be credible and persuasive. There must be at least some common understanding. If the basic values are too discrepant, the evaluator has the option of not doing the study. Of course, those basic understandings are subject to prevailing conceptions of decency and justice in the society as a whole, and the evaluator has the option of drawing upon these larger social understandings.

That is not to say that the evaluator should be in total agreement with his audiences. Presumably there are areas of disagreement or there would be no need for argument. Presumably the audiences wish to learn something new or there would be no need for evaluation. But the evaluation proceeds from areas of agreement to those areas where agreement is problematic.

## QUANTITATIVE ARGUMENT

The most popular approach to evaluation is the quantitative. Some see it as the very essence of rationality and scientific method. Many good evaluation studies have resulted from it—and many bad ones. Since this approach is taught in the graduate schools and promoted in the literature, there is little need to further extoll its virtues—they are many. In this section I would like to show that even quantitative methodology is essentially argumentation and is subject to similar considerations. Properly used, it can be a valuable tool of analysis; improperly used, it is dangerous.

Quantitative methodology is a body of mathematical methods and measurement techniques available to the evaluator. The utility of the methodology depends on similarities between the theoretical problems dealt with by the methodology and the substantive problems dealt with by the evaluator in the local setting. For his part, Cronbach (1974) has already determined that the fit of the theoretical and substantive problems is not a good one. The educational context is too complex.

In a probing analysis, a Rand Corporation mathematician (Strauch, 1976) examined the difficulties of quantitative methodology as it applies

to policy studies, i.e., questions arising from the government decision-making process. According to Strauch, in so far as the methodology is mathematical, it is a self-contained system the structure of which is determined by the premises defining the system. Mathematical analysis is the exploration of that structure as it follows logically from the premises. The results are connected to the premises by logical inference. In the sense that their validity can be determined on the basis of that chain of reasoning, the results are "objective"—there is no need to appeal to the competence or judgment of the person who produced them nor to the audience to whom they are directed. The results are necessarily logical. In argumentation, by contrast, the results cannot be totally separated from the person who arrives at them.

The application of quantitative methodology to a substantive problem uses a mathematics model as a simplified representation of the problem. The results depend in part on the mathematical analysis—but equally on the fit between the model and the substantive problem. In the simplest applications, such as in physical science, the substantive problems are rigorously quantifiable. Experimental control enhances somewhat the ability of the evaluator to make the substantive problem conform to the mathematical model, i.e., randomness in statistical models. In such cases, the conclusions are "objective" in the sense that they are subject to independent verification on the basis of the logic and fit, without reference to the judgment of the person who produced them. However, the more behavioral or political the substantive problem, the more difficult it is to define it unambiguously in mathematical terms. The links between the substance and the model become tenuous.

Strauch identifies the following components of such a quantitative study: *Formulation* involves defining the formal problem from the substantive problem, then finding a mathematical model for the formal problem. This is a process of reduction. *Analysis* involves computation within the mathematical context defined by the model. It results in mathematical statements. *Interpretation* means converting the statements back into the formal problem and finally interpreting these conclusions within the substantive context.

The validity of conclusions depends on *both* the logical validity of the analysis and the validity of the linkages. While the logical validity can be determined without reference to the subjective judgment of the analyst, the linkages cannot. They are founded upon the subjective judgments of the analyst. Both formulation and interpretation are subjective processes. Formulation requires reducing the substantive problem to something smaller that can be handled by the analysis and possibly adding some assumptions which make the analysis easier but may be questionable on substantive grounds, e.g., the independence of events.

Interpretation involves restoring the contextual considerations that have been eliminated and possibly adjusting for the simplifying assumptions.

Both formulation and interpretation require considerable doses of intuitive judgment. Hence the conclusions are not really "objective" as claimed. (See the discussion of objectivity in a later section.)

The usual way of dealing with the subjective part of the methodology is to ignore it. For one thing it is not such a great problem in the natural sciences where quantitative methods have been so successful. Evidence of "objectivity" there is taken as proof of objectivity in other areas. When these links are challenged it becomes clear enough that quite arguable premises underlie them.

Good insights are often derived from quantitative studies, but they usually result from the analyst making the right intuitive judgments rather than the right calculations. Those successes are often attributed to the quantitative methodology itself rather than to judgment. Critiques usually focus on the technical quality of the mathematical analysis rather than on the quality of judgments associated with formulation and interpretation. When quality of judgment is challenged, justification must rely on the kind of reasoning common to all argumentation.

One result of underplaying the role of judgment is what might be called "method-oriented analysis," according to Strauch. The analyst ignores the complexities of the context and plunges ahead with his favorite method. With superficial thought the methodology is applied in a straightforward manner as if there were no problems of fit. A few caveats are thrown in at the end suggesting that it is the readers' problem to decide whether the fit is a good one.

In its extreme form there is a school of thought which Strauch calls "quantificationism" which holds that quantification is a positive value in itself. A quantitative answer is always better than a qualitative one. Any problem can be reduced to a quantitative solution and no problem can be properly understood until it is. Therefore quantitative methods should be applied to all problems. This position may be a straw man in that few people would really subscribe to it.

Such an attitude, which favors scientific methodology, is based on a reductionism that treats a phenomenon as an isolated system, develops a quantitative model for that system, and uses that model as a surrogate for the phenomenon. As suggested previously, reductionism may be one element of physical science not transferable to social phenomena.

The image the quantificationist projects is of a purveyor of objective "fact" based on hard data. He takes no personal responsibility for conclusions reached by his methodology since they are not of his making. He has simply uncovered them. He is merely reporting the results of his objective methods. He disdains qualitative data as subjective.

This attitude is close to what Polanyi (1958) described as "objectivism" in science. This is an attempt to define an objective method such that it relieves the observer of any responsibility for his findings. Polanyi contended, on the contrary, that the holding of a belief requires personal

commitment and responsibility even in science. Objectivism has sought to represent scientific knowledge as impersonal.

Often quantificationism and objectivism also suit the decision maker in that he may justify his decision by reference to a "scientific" finding. It may help him avoid personal responsibility. Attempts to quantify problems that are not quantifiable and to ignore the judgmental factors eventually distorts decision making.

Strauch suggests that one way to eliminate such distortion is to use quantitative methods as a *perspective* rather than a *surrogate* for the substantive problem. Accepting the mathematical model as a valid representation of the substantive problem means using it as a surrogate. Using the model by incorporating findings into knowledge one already has means using it as a perspective.

For most substantive problems the audiences of the evaluation already have well-developed images of their own. The quantitative analysis may give the audiences an additional but not necessarily better or more valid insight into the problem. The interaction between one's own images and additional insights must take place in the heads of the audiences, the decision makers or whomever. Using quantitative methodology as only one perspective reduces the problem of the fit between the model and the problem.

On the other hand, both the evaluator and the audiences must take more personal responsibility for the findings since they do not necessarily follow from the analysis. The conclusions cannot be justified entirely on the basis that they follow logically from the assumptions. Evaluation of individual assumptions must be supplemented by holistic evaluation of the total.

Quantitative argument, then, should always be used in conjunction with human judgment, and human judgment should be given the superior position. The implications for quantitative argument in evaluation are strong. Quantitative methodology should be seen to be based on human judgments and on intuitive reasoning and should be justified accordingly.

## QUALITATIVE ARGUMENT

In his paper on qualitative knowing, Campbell (1974) indicated that scientific knowing is dependent on common sense and that particular facts from either science or common sense are known only within the body of a great many other facts. "The ratio of the doubted to the trusted is always a very small fraction." Indeed, the knowledge of any detail is context-dependent and, according to Campbell, qualitative knowing of "wholes and patterns" provides the context necessary for interpreting quantitative data. For example, generating alternative hypotheses requires familiarity with the local setting, a qualitative act.

Campbell believes that qualitative knowing has been neglected in favor

of quantitative methods. At the same time he would prefer to see qualitative and quantitative methods used together to cross-validate one another. Quantitative methods, he believes, can provide insights that the qualitative do not, in spite of the prior grounding of the latter. Also, since all knowing is essentially comparative, he thinks qualitative techniques like case studies could be improved by experimental design considerations, which he would not see as being a part of quantitative methodology.

In rethinking the necessity and even the priority of qualitative knowing, Campbell (1975) has reconsidered the "anecdotal, single-case, naturalistic observation." Quantitative generalization will contradict such knowledge at some points but only by trusting a much larger body of such observations. In the classic paper on experimental design, Campbell and Stanley (1966), the case study was described as having no basis of comparison and hence providing no justification for drawing casual inferences.

Now Campbell has modified his position considerably, coming to believe that the case worker makes many predictions on the basis of his theory which he can disconform. The process is one of "pattern-matching" in which aspects of the pattern are matched against observations of the local setting. Campbell sees the single-shot case study as being a more secure basis of knowledge than he did in the past.

How is it in Campbell's view that we can know anything? He traces the current epistemological difficulties back to a quest for certainty in knowing. The effort to "remove equivocality by founding knowledge on particulate sense data and the spirit of logical atomism point to the same search for certainty in particulars" (Campbell, 1966). Certainty was to be established by defining "incorrigible particulars." This would result in unequivocally specifiable terms and in a "certainty of communication."

Campbell now sees this brand of positivism as not being tenable in either philosophy or psychology. Things out of context are not interpretable. But how can one still "know" something from a group of events which are each in themselves indeterminate? Campbell's answer is that this is achieved through "pattern-matching."

In events of cognition like binocular vision, the eyes recognize common objects by a process of triangulation. The more elaborate the pattern the more statistically unlikely a mistaken recognition becomes. Through memory various patterns can be compared. Pattern-matching itself Campbell sees as a trial and error process. This is essentially analogical thinking and Campbell sees it as being ubiquitous in the knowing process.

In fact, scientific theory is the most distal form of knowing, and the relationship between formal theory and data is one of pattern matching with the error ascribed to the measurement of the data ("true" scores and "estimated" scores) except when it is agreed that the theory is in need of overhaul. There are two patterns to be matched, that of the theory and that of the data. Acceptance or rejection of the theory is subject to some

criterion of fit between the two. Actually a theory is never rejected on the basis of its inadequacy of fit except when there is an alternative theory to replace it. It is the absence of plausible rival hypotheses that makes a theory "correct."

Campbell sees these considerations as directly relevant to program evaluation issues. "I believe that the problems of equivocality of evidence for program effectiveness are so akin to the general problems of scientific inference that our extrapolations into recommendations about program evaluation procedures can be, with proper mutual criticism, well-grounded."

If I understand his position correctly, Campbell is arguing that evaluation is a part of scientific inquiry and subject to similar epistemological concerns. However that may be, in this paper at least, I have reversed the ground-figure relationship somewhat by treating science as an argument aimed at a universal audience and hence concerned with establishing long-term generalizations, and evaluation as an argument aimed at particular audiences dealing with context-bound issues. In any case, when two of the leading scholars of measurement and experimental design, Cronbach and Campbell, strongly support qualitative studies, that is strong endorsement indeed.

In evaluation one may think of pattern-matching occurring not only in the evaluator's mind as he constructs his study and inspects the fit between his description of the program and the actual program itself, but also in the minds of the audiences as they compare the evaluation study to their own experience. The audiences themselves have images, memories, and theories of the program under evaluation. In using the evaluation as a perspective (in this case a verbal model), the audience matches its conception of the program to the evaluation. Where it attributes the error depends on the persuasiveness of the evaluation.

The audiences thus serve as independent points of validation for the evaluation and must assume an active role in interpreting the evaluation and personal responsibility for the interpretation. In some modes of evaluation the audience may even be given explicit responsibility for approving the final report (see MacDonald's 1974 democratic evaluation in which program participants are given veto power over information about themselves).

In Campbell's terms the basic pattern-matching process is analogical rather than logical (although the process must surely involve many forms of reasoning). In fact, one can go further than this. In an epistemology based on removing equivocality and establishing certainty of knowledge by defining "incorrigible particulars," deductive and inductive reasoning are the proper way of relating these particulars. Formal logic depends on unambiguous terms operating in a closed system.

To the extent that the terms are ambiguous and the system open (or not reducible to isolated subsystems), formal logic can be applied only argumentatively. The reasoning must include other varieties of thought or one

must accept the fact that one cannot do rational analysis. Rational analysis is possible in evaluation but only rarely will it assume syllogistic form.

## AMBIGUITY AND THE DEVELOPMENT OF AN ARGUMENT

In a sense ambiguity is an essential part of such reasoning processes. Analogical or metaphorical use of concepts in evaluation will tend to render the concepts more obscure. With Campbell, I would agree that analogical thinking is basic to some forms of evaluation, and that ambiguity is a vital element in communicating experience. "Naturalistic" evaluation, for example, depends on being sufficiently ambiguous to encompass past and future cases.

In fact, some philosophers would find even pattern-matching of theory and fact as being too positivistic (Petrie, 1976). This is because observational categories themselves are believed to be determined by the theory. Without an independent observational base, there is no "objectivity" by which to assess the theory. One way around this problem, according to Petrie, is through metaphorical assertions.

The theory must prove itself against judgments of particulars. Metaphorical assertions can bridge the gap between two separate frames of reference by "showing" new relationships rather than by merely describing them. Just as a teacher uses metaphors to link what the student knows to what he does not know, scientists can explore new areas of interest by such reasoning. Petrie points to Kuhn's "exemplars" as concrete examples in science that have cognitive functions prior to specification of criteria or rules for which the exemplars are illustrations. Kuhn (1970) contended that science is actually transmitted by those exemplars rather than by idealized rules of procedure. Similarly, Petrie sees metaphor playing an essential cognitive role in both scientific investigation and in learning. Thus, qualitative evaluation may be rendered in explicit propositions, similar to scientific theses but supported by qualitative data and reasoning; or qualitative evaluation may be manifested in implicit examples of naturalistic style.

Conceptually the evaluative argument proceeds from the premises of agreement shared by the audiences and evaluator towards the perspective the evaluator wants the audiences to have. For each audience there are sets of things that are admitted and any of these is likely to affect its reactions. For example, an audience of educational psychologists will share knowledge of a set of studies (and exemplars) likely to affect their judgment of both the educational program and the evaluation. Those studies are not shared by classroom teachers. The teachers do, however, share direct classroom experiences.

The evaluator is faced with choosing themes and methods to advance the argument and appeal to the audiences. To the degree he sees evaluation as part of social science, he will use social science methodology. By selecting some elements and presenting them to his audiences, he chooses

what is important and relevant to the evaluation. He endows these elements with "presence."

"Presence acts directly on our sensibility" (Perelman & Olbrechts-Tyteca, 1969). The elements that are present to the consciousness assume an importance underestimated by more rationalistic conceptions of reasoning. The evaluator must make verbally present what he considers important, and by doing so he enhances its value. This means that the evaluative argument is inevitably selective in its presentation and is therefore always open to charges of incompleteness and partiality. The scope of the study can be enlarged but can never be complete in its coverage nor complete enough to refute the charge that something has been left out.

Partiality also exists within the large scope given to interpretation. The essential ambiguity of what things mean, even in hard data studies, causes numerous interpretation problems. Of course, the evaluator may choose to portray the ambiguity of the situation rather than to impose particular interpretations. Some British evaluators have pursued this idea most fully by refusing to provide conclusions within their evaluation reports (McDonald & Walker, 1974; Parlett & Hamilton, 1972).[4] They contend that it is the audiences' responsibility and privilege to interpret the study since only they will know what it means for them. The audiences must draw inferences for themselves based on their own experiences. The evaluator cannot be so presumptive.

British thought has long been known for its affinity for the obscure and ambiguous. As Madariaga (1949) has written in comparing the English to the French and Spanish: "The sense of the complexity of life which tends to make English thought concrete, tends to make it also vague."

Of course, it must be said that this concreteness and vagueness of thought which respects life's complexities is exercised within a strong system of traditions and roles inherent in British society. The cautious nature of public pronouncements and documents, including evaluations, is often accompanied by extreme personal opinions about the same events and personalities. In any case, the British have traditionally led the way in their appreciation of ambiguity and vagueness.

An idea is unambiguous only in a formal system in which every unforeseen element has been excluded or in which the field of application has been determined. One must be able to foresee all future cases. In such a formal system, reasoning by calculation, e.g., in chess, is appropriate. By contrast, in law a judge must make decisions that will affect future cases he cannot possibly foresee.

Ambiguous ideas can be clarified by enumerating instances but the ambiguity cannot be eliminated in this way. The context in which the idea

[4]Barry MacDonald has expressed the ultimate view in declaring that the more fully one studies a situation, the more ambiguous it will appear. If true, this raises questions about the role of evaluation in decision making, although one might contend that evaluation will only make decisions better, not necessarily easier.

is used primarily determines its meaning and a new context will shift the meaning somewhat. Analogical and metaphoric thinking apply terms to areas beyond their normal context and in a sense create new unspecified meanings. The elasticity of terms and ideas used in an evaluation means that the ideas themselves may develop and be transformed within the argument itself.

The premises will often remain implicit in an evaluative argument. There is not time to make explicit all the agreements on which the dialogue depends. All this ambiguity in choice of premises, selection of data, interpretation of meaning, and use of vague notions makes the argument nonbinding.

Indeterminacy and unspecifiability are essential parts of evaluation, whether based on hard data or soft. This ambiguity necessitates personal judgments on the part of both evaluator and audiences. It also suggests that overall judgment is more important than precise calculation in most evaluative reasoning.

## Chapter II

# THE LOGIC OF THE ARGUMENT

## MODES OF REASONING

No doubt there are circumstances in evaluation where formal logic is applicable. For example, deductive logic is certainly appropriate in determining the internal consistency of mathematical models and inductive logic is indicated in problems of statistical inference. Where appropriate, this reasoning should be applied. For the most part, however, evaluators must rely on extra-formal modes of reasoning. I will enumerate some of these techniques of argument based on Perelman and Olbrechts-Tyteca's treatise (1969) on argumentation. The list is by no means exhaustive of man's informal reasoning powers. In the next section I shall illustrate the use of these arguments by an analysis of a well-accepted evaluation study.

The techniques of argument presented here are divided into three types: quasi-logical arguments, arguments based on the structure of reality, and arguments establishing the structure of reality. The first of these types, quasi-logical arguments, derive their credibility from their similarity to formal logic or mathematical reasoning. However, it is only by a reduction that the quasi-logical argument appears to be formal. The argument is essentially non-formal rather than formal and must ultimately be defended by resort to other forms of argument.

### Quasi-logical Arguments

The first of these arguments depend on their similarity to logical relationships. They include contradiction and incompatibility; identity and definition; transitivity; and reciprocity. The other group of quasi-logical arguments depend on their similarity to mathematical reasoning. These are inclusion of the part into the whole; division of whole into parts; comparison; and arguments of probability.

*Incompatibility.* In a logical system two theses that contradict one another show the system is logically inconsistent. The quasi-logical analogue is incompatibility in which one is forced to choose between two theses that are not logically but are practically incompatible because of circumstances. In extreme cases holding incompatible theses may invite ridicule, the argumentative equivalent of logical absurdity. For example, in an evaluation the director of the project may present one view of the project while a teacher working in it may present quite a different view. The two viewpoints are not logically contradictory since both may be true as viewed from different circumstances. Nonetheless, the incompatibility may be an important point in the total evaluation. In fact, the director

whose view is incompatible with the views of others in the project does begin to look ridiculous.

*Total identity and definition.* Insofar as definitions can be stated unambiguously and unequivocally, they belong to systems of formal logic. As soon as they are applied to real world problems, definitions become quasi-logical. One must choose among many possible meanings. Only purely conventional systems can escape these identity problems. For example, validity is defined in at least five different ways ranging from a general justification to the ability to predict one event from another. One can employ any one of the definitions but the choice must be defended as appropriate and applicable if challenged by someone.

*Partial identity.* The "rule of formal justice" requires that identical treatment be given to beings or situations of the same kind. This provides for consistency of action, the basis of formal justice. "Reciprocity" of behavior rests on defining situations as symmetrical. These arguments require partial reductions, such as in the prestige and status of the parties involved, which of course depend on argued positions. For example, "It was only fair that the teacher provide special assistance to the child since she had already given extra help to others." More arguable would be "They deserved equal grades since they had exerted the same effort, although with far different results." These statements rest on definitions of partial identities.

*Transitivity.* A is greater than B, and B is greater than C, so therefore A is greater than C—but the basis of "greater than" is arguable. For example, Program A is better than B because test scores are higher. A must be better than C because B's test scores are better than C's. Of course, the criteria for comparisons are arguable as is the transitivity of the relationship itself. Program A may not be better than C even if the first relationship holds.

The arguments based on similarity to mathematical reasoning include the following:

*Inclusion of the part in the whole.* The whole is greater than each part. For example, "Having a higher total test score is better than a high score on one of the parts because the total score includes the parts."

*Division of the whole into the parts.* Exhaustive division into parts leads to the conclusion that the part left is necessary in some way. "I will list my biases for the study and against it." "Either we have a Type I error or a Type II error."

*Comparison.* Direct comparison of objects is based on an idea of measure but any standard of measurement is lacking. Criteria are often cited. Choice always implies comparison. "Argument by sacrifice" is a form of comparison: what sacrifice would one be willing to make to achieve an end? Perhaps all evaluation is basically comparative.

*Probabilities.* Argument by probability and variability usually entails a reduction of data to monistic and homogenous values and to elements by which they can be compared. But it is usually powerful because it imparts an empirical character even when non-quantitative—e.g., Decision Theory, which requires that the decision situation be reduced to a particular decision model.

Arguments Based on the Structure of Reality

An entirely different class of arguments is based on the "structure of reality." Reality is sufficiently agreed upon and unquestioned, like facts and truths, so that one tries to establish a connection between accepted notions and those being promoted. These arguments can be more finely classified as relations of succession, which relate a phenomenon to its causes or consequences; and relations of coexistence, which relate an "essence" to its manifestations, e.g., a person to his actions. Among the sequential relations, in which time plays a major factor, are these:

*Causality.* Demonstrating causal links may be based on many different methods and obviously plays an essential role in evaluative argument. The attempt to establish a causal link may involve establishing a relationship between two successive events, reasoning from a given event to a presumed cause, or projecting a causal consequence as the result of an event. In any case the causal statement requires certain value judgments (see Ennis, 1972).

*Pragmatism.* An event is evaluated by its consequences. Value of the consequences is transferred to the cause. The value of the consequences must be agreed upon or one must resort to other arguments to establish their value.

*Ends and means.* Determination of the best means depends on exact definition and agreement on the end pursued. Only values relating to the end are likely to be discussed. In a technologically-oriented society, ends-means arguments are particularly potent. Example: Behavioral objectives-programs are good which achieve these ends. Separating means and ends allows maximum agreement by separating the ends and means analytically, although it is doubtful if a particular means accomplishes only one effect. Practically, ends and means are more closely entwined.

*Waste.* Since such an effort has been exerted to this point, it would be a waste to give up now. "It would be a shame not to reanalyze this data since it has been so costly to collect." "Develop the child's talent to the fullest."

*Direction.* If we give in this time, where will it lead? The domino theory. "Knowledge can be indefinitely increased. There is no limit to learning."

*Unlimited development.* More is better and can be obtained.

Whereas sequential relations are on the same phenomenological level, relations of coexistence connect two objects or events in which one is more basic and explanatory of the other. The order of events is of secondary importance. These include the following:

*The person and his acts.* Our conception of a person is usually influenced by his actions, though ordinarily the two are not equated as they are in behaviorism. Interpreting an event by ascribing it to the personality is common practice in evaluation studies. How the "intention" of the person is handled is particularly critical. The intent is often inferred by correspondence among actions. But there is always ambiguity. Most attributions of motivation are examples of this type of argument.

*Authority.* Although rightfully excluded from demonstrations in logic, since the logic must stand on its own, the prestige of the person making an assertion is important in argument. It is essential in legal reasoning. Only if the assertion is agreed upon by the universal audience and hence considered a "fact" is it beyond the reach of authority.

"Objectivity" is often achieved by separating the person from his act, e.g., taking the author's name off proposals before judging it. However, the person may be the best predictor of the success of the project. Impartiality may be sought by bias reduction techniques rather than through complete severance of the agent from his act (see Scriven, 1975). In argumentation and evaluation the relation between a person and his assertion is important.

*Person and group.* "He did that because he's a behaviorist." This category includes arguments expressing concern in maintaining or establishing relations with others. Characterizing a person through his group membership is far more common in evaluation than is realized. Not only are quantitative studies set up to reveal differences among groups, qualitative evaluations often interpret the social system under study as a set of interacting groups. In addition, the evaluator is often at pains to demonstrate his concern and/or impartiality by showing what groups he himself does or does not belong to.

*Acts and essence.* What is a good director? A good director is one who conforms to the ideal of a director. In the absence of such conformity there is a "deficiency." The essence of an object under evaluation is often defined by a set of intuitive criteria one would expect to apply. For example, a "good project director" would be expected to be and to do certain things. The evaluator may elicit this normally implicit set of criteria in order to judge the director. The same thing can be done with a good program, a good textbook, etc. The list of criteria is never inclusive and is always arguable. Nonetheless, the list is often effective in persuading the audience as to quality. Example: Consumers Union reports on manufactured products.

*Symbolic relation.* Only members of a particular group believe in the magical relationship between the symbol and the thing, such as a national flag. Symbolic relationships are important in describing certain aspects of social systems and statuses. These relations are somewhat different in that they cannot be justified to others. Educators often attach such special meanings to particular facets of their program and to particular charismatic leaders within it. People and things become the objects of faith in and of themselves. This is a common puzzle to the evaluator who may look in vain for more material relationships underlying the faith.

Arguments Establishing the Structure of Reality

The third class of arguments assumes the fewest premises in advance. These arguments rely neither upon similarity to formal logic nor argue from the already agreed upon structure of reality. Rather they try to *establish* reality. Example and illustrations do so by resorting to the particular case. Analogies and metaphors do so by showing new conceptual relationships to the audiences. This mode of argument is relied upon heavily in "naturalistic" evaluation.

*Example.* Resort to example implies lack of agreement on a particular rule but a prior agreement that one might eventually come to an understanding. A series of examples induces one to generalize. Sometimes the reasoning is from the particular to the particular with no rule being stated. The examples operate implicitly. The technique of the "closed case" and the legal "precedent" is built on such a technique. This argument values the actual and the habitual. To be effective the example itself must be accepted as factual.

*Illustration.* Whereas example is used to establish a rule, illustration is used to clarify one and strengthen adherence to it. It promotes understanding. Illustrations of forms of arguments in this section attempt to clarify the categories but the categories are not dependent on the illustrations.

*Analogy.* Analogy strikes a relation between two previously unrelated spheres and is hence essential in invention and imagination. It develops and extends thought.

*Metaphor.* Metaphorical assertion opens new realms of thought by moving from the known to the unknown and by helping indicate things unspecifiable in ordinary language. Metaphoric assertion is most used in conjunction with examples and illustrations. How it works to extend the audience's ideas will be discussed as part of naturalistic evaluation.

These techniques of argument are not exhaustive and are not intended as a list of techniques from which to construct evaluations. Rather they are meant to illustrate the kind of reasoning that is actually employed in evaluations.

## ANALYSIS OF GLASS'S "EDUCATIONAL PRODUCT EVALUATION"[3]

I have chosen Glass's "Educational Product Evaluation: A Prototype Format Applied" (Glass, 1972) to analyze in terms of the arguments enumerated in the last section. I selected this evaluation for several reasons:

1. It is highly accessible, having appeared in the *Educational Researcher.*
2. It is a succinct evaluation.
3. The authority of the author is unassailable.
4. It exhibits a variety and complexity of evaluative arguments.
5. I find it personally quite persuasive.

My technique will be to paraphrase Glass's work and to identify the arguments in parentheses as they occur. I would not contend that I have found all the arguments in Glass's work, that the ones I have emphasized could not be categorized otherwise, or that the types of argument I have enumerated in the last section are exhaustive. It would be impossible to list all arguments or types or to classify them unambiguously. My purpose is to illustrate from a very good piece of work that those arguments play a critical role in evaluative reasoning. The overall logic of the Glass piece is somewhat more complex than the arguments I have discussed, and I will save it until after a discussion of particulars.

Glass begins with a brief introduction stating the tentative nature of evaluation techniques and describing what he intends to do. The body of the paper is divided into ten parts. Part I is a description of the AERA cassette recording he intends to evaluate, which is itself a discussion of evaluation by Michael Scriven.

[3]Glass's evaluation has been reproduced in the appendix.

Part II lists the three goals of the product and evaluates them. Training evaluators is good since there is a need for evaluation skills because of legislation mandating evaluation (cause and effect). Producing a cassette that can be used while commuting to work may or may not be desirable because it may infringe upon a person's private time in unanticipated ways (pragmatic argument—valuing an event from consequences). Experimenting with new media is commendable if it is not "mere technological tinkering" (person and his actions—intention of the actor). The evidence will be whether the cassette is properly evaluated (intention constructed from consistency of actions—person and his actions).

Part III describes where things stood as the evaluator entered. The director, the topics of the tape, the lecturer, the subject matter, and the initial copies have already been agreed upon. The vending of the cassettes, the choice of materials, and marketing plans are not settled. This signals where it is reasonable for Glass to focus attention. Implicit is the argument that it would be a waste of the evaluator's and audiences' time and effort to address issues already decided (argument of waste).

Part IV is entitled "trade offs" and is a brilliant turn in the overall argument. Glass enumerates what could be purchased with the resources used to produce the cassette—one day of training session for 100 researchers, printing of 20,000 copies of prose materials, a half-year stipend for a research trainee, or four scholarships to AERA training sessions for minority researchers. This is the trade-off for the sponsor, the USOE. Trade-offs for the other major audiences—the director, AERA, and the consumer—are also listed.

The reasoning begins by asking what would be given up by the cassette approach (argument by sacrifice). It establishes the equivalence of the trade-offs in terms of their being purchasable with the resources devoted to the cassette approach (argument by identity). The trade-offs are also equivalent in that they are all consistent with the producer's intent.

Without making explicit the reasons, Glass chooses the typescript alternative as the trade-off "with the greatest leverage" (argument by comparison). Why choose the strongest alternative with which to make further comparisons? Implicit in the reasoning is the idea that one should choose the technique which will *best* further the end of the producer (argument by ends and means).

Having chosen the strongest competitor, Glass, in Part V of the study expands the cost comparison between the cassette and typeset approaches to the fullest (arguments by comparison and sacrifice). In exploring cost considerations, he argues that the cost would be worthwhile for groups of 10–15; that the tape is too expensive and could be cheaper—for this he cites the Colorado audio-visual instruction department as authority (argument by authority); that typescripts could be better stored; and that the typescript's cost could be further reduced. All these arguments are varia-

tions based on comparisons between the two approaches and what each might cost under various contingencies.

Part VI is the "intrinsic" evaluation, labeled secondary by the evaluator. It is an evaluation of the technical quality, content, and "utilization of uniqueness" of the medium. This series of arguments deals with issues that are secondary to the entire cassette versus typescript comparison but which might be important to a potential consumer who wishes to purchase the cassettes.

The evaluation of the technical quality and content are based on an ideal of what the technical quality and content should be—deviations from these ideals are deficiencies (argument by act and essence). The evaluator lists criteria which he considers to be relevant and commonly agreed upon, since he does not attempt to justify them. Technical quality contains tape quality, recording fidelity, aesthetic quality, editing, and packaging. Each criterion is accompanied by a judgment and a few remarks enumerating observations on which the judgment is based. Similar "a posteriori" criteria are applied to the content.

The second part of the intrinsic evaluation is of the "utilization of uniqueness" of the cassette medium. This is again basically an argument based on the "essence" of the cassette (act and essence). Two producer claims are explored. The fact that one can stop the tape advantageously is refuted by the evaluator by counting the number of stops. The second claim that a significant number of people have cassette players and time in which to listen to the cassettes is confirmed by a mail survey to 100 AERA members (argument by probability). Knowing he is addressing an audience of educational researchers, Glass reports the confidence intervals in a footnote. Throughout the second part the dormant comparison with typescript is utilized by refuting producer claims that reading typescript cannot do the same things. Glass argues against the producers' "unique features" claim for the cassette approach (argument by act and essence).

Part VII is the "outcome" evaluation and is labeled as primary by the evaluator. The comparison between cassette and typescript is head-on in terms of outcomes. He argues that even if the aural medium is as effective in transmitting information, it is slower. This is a comparison implying measurement. It is a comparison based on pragmatic consequences (argument by comparison; pragmatic argument). Access is also much slower on the cassette (argument by sacrifice).

Glass cites a review of experimental studies comparing the aural versus visual mode as being inconclusive because relative efficiency depends on several contingencies. This is non-contributory to his argument, other than increasing the evaluator's credibility, but it allows Glass to describe a particular study in detail which shows the superiority of visual learning (argument by illustration).

Part VIII is a summary of conclusions and a separate set of recommendations for each separate major audience. The recommendations are quite

direct, explicit, and specific in their direction. In fact, the recommendations establish a hierarchy of actions each audience might take, depending on contingencies. Part IX lists the special audiences who might benefit from the cassette approach. The arguments are that cassettes may be beneficial to sightless learners, large groups, "Reverse Luddites." All these arguments in Parts VIII and IX are variations of costs and benefits (arguments of ends and means; pragmatic arguments).

Part X is unusual in its reflexiveness. It is entitled "Evaluating the Evaluator" and explores the evaluator's own biases. Of course, simply undertaking such a consideration enhances the evaluator's credibility. Glass points out that evaluations themselves involve costs, especially in destroying a sense of community (arguments of person and group). In this case, he undertook the study because he was asked by the product developer (person and act). He establishes his credibility by showing that he took actions which are inimical to his own interests, thus giving evidence of his impartiality.

Glass divides his motives into the exclusive categories of motives for a favorable evaluation and motives for an unfavorable evaluation (argument by division of whole into parts). Biases for a favorable review derive from the fact that Glass is a member of the AERA Executive Board, the benefactors; and the fact that the producers are his close colleagues (argument of the person and his group).

Motives for the unfavorable review are that he declined to participate himself on the grounds the cassette approach is not cost effective and the fact that he was once beaten in table tennis by the project director. These arguments depend on the construct of the person behind the acts (argument from person and acts). He concludes the evaluation by pointing out that he has collected no data on attitudes toward the product or on its effectiveness. He leaves the audiences to draw their own conclusions on the balance of biases and overall credibility.

The *overall* structure of the study is well worth examining. It consists of a complex form of argument called the "double hierarchy" argument (Perelman & Olbrechts-Tyteca, 1969). The double-hierarchy argument consists of two hierarchies of values or objects which are usually connected by relations from the structure of reality. For example, Leibniz' statement that "since [God] cares for the sparrows, he will not neglect reasonable creatures who are far dearer to him" is based on implicit hierarchies of creatures and God's caring and connected by implied cause and effect. Double-hierarchy arguments often take the forms of "If . . . then" conditional statements and are usually implicit.

The overall logical structure of Glass's evaluation seems to consist of a double hierarchy argument. One hierarchy is a hierarchy of costs. The other hierarchy is one of benefits. The two hierarchies are connected by a means-ends relationship. In fact, the entire study is based on establish-

ing this logical structure and orchestrating the subarguments within the grand overall design.

For example, after the context of the study is defined by the product description, the producers' goals, and the entry point of the evaluator, Glass builds a hierarchy of trade-offs in Part IV. In Part V he selects the strongest competitor and builds the cost comparison hierarchy between the two approaches. In Part VII he builds the benefits hierarchy, again based on comparisons between the two approaches. The means-ends relation connects the two hierarchies. It demands that the best means be chosen to accomplish given ends. The contingencies in Parts VIII and IX are explorations of what would happen if one moved up or down the cost hierarchy or the benefits hierarchy.

Thus Glass has conducted a cost-benefit analysis without precise measurement of the costs or the benefits. And it is persuasive. It is so compelling, I think, because of the integration of the arguments. All the arguments work economically within the overall structure. There is very little extraneous movement. Only the introduction and the final section on the credibility of the evaluator do not contribute directly to the overall argumentative line. Aesthetically these two sections are appropriately placed at the beginning and end. One is inclined to agree with Polanyi that the ultimate test of truth is the coherence and beauty of the structure.

The most difficult part to handle in the overall design is Part VI, dealing with the quality of the cassette. Glass was actually asked to evaluate the cassette itself. I would surmise that the basic problem of intellectual incompatibility from which the evaluation grew was that the cassette itself was good but Glass did not see the investment as being worthwhile. He redefined the problem such that he was evaluating the cassette approach rather than just the cassette itself. Yet he could hardly evaluate the product without direct evaluation of the tape. Also, one of his audiences had to be potential consumers who might buy the tape and not just AERA board members who wanted to know if the entire activity was worthwhile. He labeled the cassette evaluation secondary as opposed to the primary outcome evaluation. Aesthetically he also de-emphasized it by tucking the intrinsic evaluation into the middle of the overall presentation.

In addition to the logical coherence of the evaluation, it is also persuasive because the premises of agreement are well chosen for the audiences. Costs/benefits are powerful values for the audiences and means-ends relations are nearly unquestioned by people versed in the teleology of utilitarian ethics. Glass takes the audiences from values they agree with to conclusions they may not have accepted initially. He is keenly aware of who his audiences are, even addressing each directly and giving each different recommendations. One may suspect, however, that his arguments are not equally persuasive to all. Some groups are likely to harbor values and conditions untouched by the evaluation. Yet he has solved the

problem of composite audiences with differing demands beautifully, both logically and aesthetically.

How would one deny such an evaluation? One could attack the basic argumentative structure by denying the equivalence of the trade-offs and by questioning the selection of the strongest competitor, thus denying the means-end relationship. One could attack the costs and deny the comparative benefits that result from the typescript approach. Attacking the secondary evaluation of the tape quality itself does little good since it is not integral to the overall logic of the study. Glass can concede points there and still arrive at negative conclusions. One can also claim the evaluator is unduly biased and attack the credibility of the study in that way, although Glass's discussion of his own biases makes it more difficult to do. Any evaluation is assailable, even one that is highly persuasive.

It is noteworthy that in this masterful evaluation, Glass has used most of the types of argument previously enumerated. He relies heavily on arguments from the structure of reality, especially sequential relationships linking phenomena to consequences such as ends and means arguments, and on quasi-logical arguments such as comparisons. He has very few arguments which attempt to establish the structure of reality such as examples and metaphors.

Formal data collection procedures are used only moderately, and where employed do not contribute critically to the import of the evaluation. Most data consist of already accepted "facts." Formal data collection procedures are not essential to evaluation; argumentation is.

## ANALYSIS OF SCRIVEN'S[6] RESPONSE TO GLASS'S EVALUATION

This section was written five months after the rest of the paper because I did not know of Scriven's response to Glass's evaluation until informed of it by Glass. The timing is important because Scriven attacked Glass's evaluation in precisely the way it was suggested in the previous section one would have to do. One must deny the equivalence of the trade-offs and question the selection of the strongest competitor, thus denying the means-end relationship, as well as attack the costs and deny the comparative benefits of the typescript alternative. This is what Scriven does, and comparing his reasoning to Glass's is interesting.

Scriven (May, 1972) begins by saying he has been invited to respond to the Glass evaluation of his cassette (intentions of the actor—argument relating a person and his acts). He sketches the background conditions surrounding his decision to redo the entire second cycle rather than revise the first product. The argument is laid out rationally as a choice among three alternatives (pragmatic argument). However, Scriven devotes so much space to developing the context of his action that he clearly wants

[6]Scriven's response has been reproduced in the appendix.

his audiences to understand his motivations (argument relating a person to his acts).

Scriven also has a much larger problem with bias than does Glass because Scriven is responding to an evaluation of his own product and is immediately suspect. Interestingly he argues that the direction of bias is so obvious that it can do no harm (pragmatic argument). This argument attempts to *sever* the relation between the act of counter argument and the motivation (bias) of the actor (argument relating a person and his acts). It is an attempt to reduce perceived bias on the part of the actor. Scriben buttresses his impartiality by showing his ability to distinguish between "excuses" and "criticisms" (in itself an argument by division of whole into parts). While there are several kinds of arguments in the first two sections of his response, Scriven organizes them towards disassociating himself from bias.

In the third section Scriven turns to a consideration of the conclusions about the hardware. He accepts most of Glass's criticisms (enhancing his credibility) but dismisses the desirability of cheap tapes because they will wear badly (pragmatic argument) because of heating and friction effects (cause and effect). He also dismisses distortion in the cassettes if they are played on the proper equipment. The mention of Advent and MacIntosh equipment immediately captures the audiophiles in the audience and shows that Scriven knows what he is talking about (argument by authority). Now it is clear why he started with an analysis of the relatively unimportant area of hardware; Scriven has better information in this area than does Glass. It is also an attack on Glass's cost analysis in terms of the size of the audience reachable and the cost of the tapes.

There is little to argue about in software since Glass's evaluation of Scriven's tape is a string of "excellences." Scriven dismisses the criticism that lack of citations is a handicap, based on the feedback he has received from the field (argument by probability).

Then comes Scriven's basic attack on the logic of Glass's evaluation. Scriven concedes that "the general procedure of really working to get estimates of comparative cost-effectiveness seems to me absolutely correct and indeed the method of choice in all educational evaluation." But he is not in agreement with Glass's assessments of the costs and benefits and particularly the way Glass has them linked together. Scriven's basic thrust is that Glass has chosen the wrong competitor (the typescript) for comparison.

Scriven contends the cassette serves different ends than does the typescript; it is more useful than listening to a car radio and it can be a cheap surrogate for a visiting lecturer in a course. (These two exclusive purposes are established by definition.) These arguments deny the equivalence of outcomes that Glass has established (the argument by identity). The cassettes accomplish different ends and therefore the trade-offs are not equivalent. The cassette is a motivator in places where written material

is not (pragmatic argument). Also the costs are the same as for commercial tapes (comparison with the norm).

Scriven admits cost, speed, and replay advantages for the typescript, but again the cassette introduces a new element the written material does not. Scriven gives several reasons for using the cassette in class: hearing the authority himself; several speakers are better than one; and the tape provides variety (arguments of pragmatism, the whole greater than its parts, and unlimited development). While not generally superior, it is "repertoire-enlarging." Notice that overall, Scriven is arguing for the *uniqueness* of the cassette while Glass is arguing that the typescript accomplishes more of the common goals (loci of quality versus loci of quantity).

The cost trade-offs Scriven treats as problematic. Perhaps the funds for producing the cassettes were not available for anything else under the circumstances (argument of waste?). Even if they were, AERA should be doing experimental things (act and essence—"being experimental" an implicit criterion for AERA); and this is a reasonable experiment, given other attempts (argument by comparison with the norm). Also it is better to try it in education if it is to be used in education (partial identity?).

But Scriven's main objection to Glass's evaluation is the object of comparison: "So my principal criticism of the Glass evaluation concerns the choice of the main crucial comparison. It should not have been the typescript but just the better content—cheaper package cassette." The disagreement is not merely one of comparison. The disagreement is whether to connect the costs and benefits by a means-ends argument, which suggests the *best* competitor—the typescript—or by a pragmatic argument, which suggests a *lesser* competitor.

Scriven insists on the uniqueness of the medium. Although Glass has refuted the uniqueness argument by counting the number of times Scriven stopped the tape, Scriven argues he is not persuaded because Glass did not offer what would be a unique utilization. Scriven switches to "comprehensibility" as the uniqueness factor, admitting that the number of stops on the cassette is a poor indicator of that criterion (argument by act and essence).

In the last section of his response, Scriven suggests that Glass's "Reverse Luddites" category of potential audiences is too narrowly conceived and that there are many normal people who would benefit from a cassette because there are people who prefer listening to reading (arguments by frequency). In fact, everyone does so at some time of the day (cause and effect). These arguments are supported by Scriven asking his wife (argument by illustration) just as Glass used a study by one of his graduate students. Finally Scriven says one must also consider the additional benefits of what has been learned by the intermediary population—himself and the producer (pragmatic argument). All these arguments increase the benefits, thus making the cost benefit ratio more acceptable.

The overall logic of Glass's original evaluation is a double hierarchy argument of costs and benefits linked by a means-ends relationship. Scriven sees this structure clearly and accepts the basic comparison of costs and benefits as the method of choice for all evaluations. He tries to show how the costs are not extravagant and unreasonable and that the benefits of the cassette are significantly underrated by Glass. But the main criticism is to challenge Glass's means-ends argument by substituting a pragmatic argument as the link. The means-ends argument requires that the cassette be compared to the *best* alternative available. Scriven's pragmatic argument requires only that the cassette be better than what now exists among other cassettes. Scriven's strategy is to claim unique features for the cassette so it does not have to compete totally head-to-head with the typescript approach on each dimension. Scriven is arguing for a qualitatively different field of comparison.

The pragmatic argument in its elemental form consists of evaluating an event in terms of its consequences. The means-ends argument, on the other hand, depends on agreement on the ends. Determining the *best* means to the ends depends on *exact* definition of the ends pursued. Values not related to the ends are eliminated from consideration. If the ends are exactly defined and agreed upon, the determination of the best means becomes a technical problem. Such reasoning, appropriate for the technical disciplines, is quite different from every day reasoning.

Generally speaking Glass's work as a whole has tended to be more means-ends and more technically oriented while Scriven's has tended to rely more on pragmatic argument. In fact, Scriven's goal-free evaluation might be regarded as an ultimate expression of pragmatic argument. One does not care about the expressed ends at all but only about the consequences of the object under evaluation. Generally, conceiving an evaluative problem in "means-ends" logic tends to devalue the means in relation to the ends, while conceiving the same problem in "event-consequences" logic tends to make the event relatively more important. Scriven's challenge to Glass culminates eventually in a discussion over the ends of the cassette approach.

On a more abstract level the dispute is between two principles of rational choice: the principle of effective means and the principle of inclusiveness (Rawls, 1971). The principle of effective means stipulates that, given the objective, one is to achieve it with the least expenditure of means or, given the means, one is to fulfill the objective to the fullest possible extent. In other words one is to adopt the best alternatives.

The principle of inclusiveness stipulates that one alternative plan is to be preferred to the other if it would accomplish all the aims of the other plan plus some additional aims. In arguing for the cassette approach as "repertoire" expanding but not as a total substitute for the typescript, Scriven is so arguing.

The few differences between Glass and Scriven should not obscure the

many similarities of their evaluative argument. Both accept comparison of costs and benefits as the method of choice. Both rely heavily on "structure of reality" arguments, Glass relying a little more on relations of coexistence, e.g., the relations between a person and his acts and between a person and his group. Scriven relies slightly more on sequential relations arguments, especially pragmatic argument. In spite of structure of reality arguments, there is little surveying of others for information. Both rely on their own personal observations for primary data.

Secondarily, both use quasi-logical arguments, though only about half as often as the above arguments. Both use arguments attempting to establish the structure of reality, e.g., examples, analogies, etc., only once. An entirely different type of evaluation would have been to put the cassettes into use in the field and to collect anecdotes about how they are used. This type of evaluation will be discussed in the next section as "naturalistic" evaluation.

Both Glass and Scriven use more than twenty-five arguments in their articles, although Scriven's article is half as long as Glass's. Scriven's high argument density reflects his general style: he is apt to spin out a number of reasons for a given judgment one after the other in a profuse and linear fashion. Here and elsewhere, Glass offers fewer reasons but they are more carefully articulated with one another, some arguments carefully nested within others.

Partly because of this, Glass's piece is more coherent and aesthetically pleasing than is Scriven's. Scriven is at the disadvantage of having to respond to Glass's paper rather than creating a full-fledged argument form of his own, as he did, for example, in his goal free evaluation paper (Scriven, 1973). The somewhat rambling flow of Scriven's response as he answers various points in Glass's paper detracts from the overall persuasiveness of his arguments. It is a serious disadvantage that every respondent to a document must face.

Finally it should be noted that this exchange between two of the foremost evaluation theorists is not primarily over data. Rather the dispute is over the proper comparison for the object under evaluation, which is eventually traceable to the argument form preferred and the audiences addressed. Some people think that all disputes can be resolved by data but such is not the case. It is often the logic of the evaluation that is in dispute.

## NATURALISTIC EVALUATION

When one reads a novel or poem, something is learned. If someone were to ask what has been learned, it would be difficult to say. Often the knowledge gained from such reading is not in propositional form. Yet in the reading of such works, experience from the novel or poem is mapped onto the mind of the reader. The kinds of generalizations the reader acquires

have been called "naturalistic" (Stake, 1976) or "spontaneous" (Perelman & Olbrechts-Tyteca, 1969).

The class of arguments that try to *establish* a structure of reality and assume the least agreement in advance between the author and audience are those most used in "naturalistic" evaluation. They include example, illustration, analogy, and metaphor. I would label as "naturalistic" an evaluation which attempts to arrive at naturalistic generalizations on the part of the audience; which is aimed at non-technical audiences like teachers or the public at large; which uses ordinary language; which is based on informal everyday reasoning; and which makes extensive use of arguments attempting to establish the structure of reality. In this category I would include most case study evaluation (Stake, 1976; Smith & Pohland, 1974; Parlett & Hamilton, 1972; and MacDonald & Walker, 1974) and also those employing legal procedures (Levine, 1973; Owens, 1973; and Wolf, 1974).

Denzin (1971) described the naturalistic approach in sociology. It attempts to blend the "covert, private features of the social act with its public, behaviorally observable counterparts. It thus works back and forth between word and deed, definition and act." The observer is a part of the research act and reflections on the self may be important data. The research begins with troubling issues and admits any and all relevant ethical data.

The focus is on the complexity of everyday life, and naturalism tries to understand the everyday world in the experience of those who live it. The naturalist shows profound respect for the empirical world. Participants serve as constant sources of ideas and as checks on the developing ideas of the naturalist. Multiple perspectives are essential to portray the whole picture. The naturalist carries on and perhaps records covert dialogues with himself as he tries to explain events.

Since the focus is on understanding various interactions, the naturalist must follow events over time. He searches for explanations, rather than predictions; and explanations must usually be grounded in the retrospective reasons people give for their own and others' behavior. This necessitates considerable submersion in the participants' culture and language. Joint actions are major points of attention, and they have to be seen in some historical perspective.

Validity is provided by cross-checking different data sources and by testing perceptions against those of participants. Issues and questions arise from the people and situations being studied rather than from the investigator's preconceptions. Concepts and indicators "derive from the subject's world of meaning and action." In constructing explanations, the naturalist looks for convergence of his data sources and develops sequential, phase-like explanations that assume no event has single causes. Working backwards from an important event is a common procedure. Introspection is a common source of data.

Of course, the sociologist is interested in constructing a generalizable theory. The naturalistic evaluator is interested only in the case he is evaluating. The sociologist will try to justify his conclusions to a universal audience. The naturalistic evaluator must adjust his work to a particular audience, who may even be the participants of the program he is evaluating. In presenting their studies both will rely heavily on examples and illustrations drawn from the field. The evaluator may or may not draw specific conclusions from the examples. If the examples are collected and presented systematically, their logic will resemble that of inductive reasoning. However, in naturalistic evaluation the audience always has the choice of how to interpret the findings and of how much credibility to assign them.

Evaluations using examples and illustrations extensively, even evaluations which consist entirely of one extended example, are becoming commonplace. They are particularly important when appealing to nontechnical audiences who are not familiar with more arcane forms of quantitative argument and to audiences for whom the evaluator can make few assumptions about the premises of agreement. School practitioners fall into both these categories. It is dangerous to presume that practitioners start from the same values and see reality the same way as evaluators or government officials.

Analogies and metaphors are seldom used in evaluation, because they are often perceived as mere figures of speech and thus unreliable data. They are, however, important ways of arriving at naturalistic generalizations. Petrie (1976) suggested that Kuhn's exemplars convey cognitive categories essential for an initiate to understand scientific theories. Ortony (1975a, 1975b) discussed the ways in which metaphors work to extend thought.

Ortony contends that words do not precisely convey the flow of experience as it is presented to the human mind. Experience is continuous and non-discrete; and even though words do not have distinct meanings like logical symbol systems, neither do they accurately represent all forms of experience. By "particularization" metaphors help bridge the gap between language and experience. Particularization conveys mental images to the mind of the reader. A term like "fearless warrior" evokes meaning more succinctly and compactly than does a longer description. In addition metaphors can capture distinctions that are otherwise inexpressible.

According to Ortony, another characteristic of metaphors is their vividness. They are closer to experience and convey emotional as well as cognitive and sensory meanings. This imagibility is associated with learnability. Metaphors facilitate insight and personal understanding by moving from the known to the less known. They facilitate naturalistic generalization on the part of the audiences. It is critical, however, that the author understand his audiences in order to know whether a metaphoric assertion will expand understanding or simply pass the audiences by.

Ortony also extends this conception of language into the teaching-learning situation. Drawing upon Polanyi's idea of tacit knowledge, he contends that the teacher must always know much more than he can express in propositional form. It is this tacit knowledge, partially a knowledge of contextual application, that is the deep understanding of a field or discipline. In order to communicate knowledge to a student, the teacher must select from his tacit knowledge and try to represent it in propositional terms. The propositional form is always somewhat removed from the full tacit understanding.

The student initially sees only the propositions. It is like learning to ride a bicycle by reading a set of instructions. The beginner's behavior is controlled by the explicit propositional knowledge which is inadequate. It is here that the teacher can aid the student by examples, metaphors, and non-literal language.

Scientists trying to learn their discipline have similar problems. According to prominent critics, it would be impossible to learn a scientific discipline by following a set of rules (Polanyi, 1958; Kuhn, 1970). According to Kuhn, a scientist learns his discipline through a set of exemplars—concrete problems permitting solutions that enable the novice to make comparisons with other disparate problems. The shared meaning is transferred through these experiences and not through rules.

The similarity between naturalistic generalizations in evaluation through the use of examples and metaphors and other arguments which attempt to establish a structure of reality is clear. Understanding and insight on the part of the audience is facilitated even though there may be no scientifically verified propositions in the sense of formal logic. Even though its epistemological and psychological assumptions are somewhat different from other types of evaluation, naturalistic evaluation is still a form of argumentation.

## OBJECTIVITY, VALIDITY, AND IMPARTIALITY RECONSIDERED

What does it mean to say that an evaluation study is "objective" or "valid?" Few concepts have been so confused and have caused so much mischief in educational inquiry. Many people are reluctant to accept or believe qualitative evaluations simply because they are based on only one person's observations. Observations by one person are considered in and of themselves to be subjective and hence illegitimate for public purposes.

The crux of the confusion lies in misconceiving "objectivity." Scriven (1972) has written cogently and brilliantly about this confusion, tracing the unfortunate history of how objectivity has been defined. The theme of most definitions of objectivity is that there is something outside the mind that is verifiable through public or intersubjective agreement and that one can express or prove such things without influence from personal feelings. An evaluation which can do so is objective. But can one person's view ever

be "objective?" The difficulty lies in confusing objectivity with procedures for determining intersubjectivity.

Scriven (1972) contends that there are two different senses in which objectivity is used—the quantitative and the qualitative. In the quantitative sense of the term, one person's opinion about something is regarded as being subjective—the disposition of one individual. Objectivity is achieved through the experiences of a number of subjects or observers. The common experiencing makes the observation public through intersubjective agreement. More formally, one might say that with a number of individuals one is more certain that one has properly represented the population—a sampling problem.

The qualitative sense of objectivity is quite different. It refers to the quality of the observation regardless of the number of people making it. Being objective means that the observation is factual, while being subjective means that the observation is biased in some way. Is it possible for one person's observations to be factual while a number of people's observations are not? Indeed it is. So an observation can be quantitatively subjective (one man's opinion) and also qualitatively objective (actually unbiased and true).

In fact, one might contend that the types of biases that affect the opinion of one person are somewhat different from those biases that plague group opinions. For example, an individual may succumb more easily to idiosyncratic viewpoints since he can hold only one perspective. On the other hand, there are social and cultural biases to which a group is more susceptible than is a particular person, e.g., jingoism. The individual's qualitative objectivity can be assessed by his previous track record on such matters and by his current self interests. In any case, one who subscribes entirely to the quantitative notion of objectivity is not going to be satisfied with approaches like case studies.

How did the quantitative notion equating the number of people making an observation with its truth gain such ascendency, even to the point of excluding qualitative objectivity? Scriven traces this distortion to psychology's attempt to root out introspectionism and philosophy's attempt to purge obscure metaphysics. Both tried to do so through the verification principle. Intersubjectivity became operationalized as *the* criterion for objectivity. In its extreme form the equating of objectivity with the quantitative notion of intersubjectivity was manifested in methodological behaviorism and in operationalism. But the fallacy of intersubjectivism pervades all fields.

Scriven cites the example of an evaluation of a television antenna in an electronics magazine in which the evaluator can see and report a better picture resulting from one of the tested antennas. Yet the evaluator apologizes for being "subjective" in his approach *since he did not use an instrument to measure decibel gain.* In fact, as Scriven notes, it is possible to get intersubjective agreement without instruments on the performance

of electronic equipment and it is the case that these pooled judgments of quality do not correlate highly with any instrument readings. Why then is an instrument reading objective while one person's judgment is subjective in the perception of this confused evaluator?

The reason is that the evaluator is only one person making the observation, and even though he knows he could have his observation confirmed by calling in his colleagues, he believes an instrument would be better because he can get even higher agreement among observers on the meter reading itself—even though the meter reading is not highly indicative of quality. In this case the quantitative notion of intersubjectivity has supplanted the quality of the perception.

In operational terms "measuring on a quantitative scale by mechanical means" becomes the indicator of truth because the interjudge reliability is higher, according to Scriven. Simultaneously one has actually sacrificed validity for reliability because the meter reading, while reliable, is not a good indicator of picture quality. This is one of the common errors of evaluation—the substitution of instruments for direct observation of quality, the substitution of reliability for validity. And it is an error of the first magnitude.

From this idea—that what cannot be directly experienced by others cannot be taken seriously by science (intersubjectivism)—has developed the concept of objectivity as the externalization of all references so that multiple witnessing can be achieved, a gross oversimplification according to Scriven. In educational inquiry, this has been manifested in equating objectivity with the ability to specify and explicate most completely all data collection procedures. Complete externalization and objectification permit replication, the hallmark of reliability. In education being objective has come to mean having a "valid" instrument—just as with the electronics evaluator.

What exists, in fact, are highly reliable instruments the validity of which is questionable. They do not always correlate highly with judgments of educational quality. The distortion of the intersubjectivist verification principle has resulted in equating objectivity with externalized, replicable procedures—even though these procedures may be infected by biases and hence be qualitatively subjective.

The identification of objectivity with a completely specifiable external procedure has another important effect. It relieves the evaluator of responsibility for the results and consequences of the evaluation. After all, if these "objective" instruments and procedures give these results, how can the evaluator be held liable? Science is to blame. Polanyi (1958) calls this position "objectivism." Objectivity in this sense comes to mean that observations are subject to independent verification without reference to the person who produced them.

Now it is not possible to specify all knowledge explicitly nor to verify it completely by independent-external procedures. Scriven contends that

even in mathematical proofs in which the steps of the proof are reduced to the self-evident, intuition plays an inevitable and important role. Not only is intersubjective verification not a guarantee of truth, it is not even necessary. Truth is an ideal which is approximated through an interplay of introspection and public verification.

Because of their complexity, many intuitive judgments can never be fully explicated. But conclusions may be no less true because of one's inability to explicate them. Agreement among many may be necessary for explaining the truth to someone else but it is not necessary for the truth itself.

How is it possible to establish the validity of a claim if one cannot separate it entirely from the person making the claim? One way is to check the reliability of the observer in previous instances and to check the observer's freedom from bias. These are not guaranteed to produce truth but there are no guarantees anyway. There are knowledge claims that are hybrids of the internal/external split, e.g., tendency statements, analogies, approximations, that are true yet are not the types of claims one usually associates with scientific statements, according to Scriven. He calls them "weak knowledge" claims and suggests they represent the type of knowledge available in the social sciences.

Such knowledge claims are manifested more as explanations than as predictions. Explanation and understanding are functions of the way information is coded in the mind. Explanation implies a person who is understanding the explanation. It does not exist by itself. The understanding is ultimately reducible to something familiar in the mind of the audience doing the understanding—or else it is not an explanation.

Similarly, unless an evaluation provides an explanation for a particular audience, and enhances the understanding of that audience by the content and form of the arguments it presents, it is not an adequate evaluation for that audience, even though the facts on which it is based are verifiable by other procedures. One indicator of the explanatory power is the degree to which the audience is persuaded. Hence an evaluation may be "true" in the conventional sense but not persuasive to a particular audience for whom it does not serve as an explanation. In the fullest sense, then, an evaluation is dependent both on the person who makes the evaluative statement and on the person who receives it.

Prediction is not necessary to demonstrate understanding. Inferring another event from a correlation coefficient plus a few antecedent conditions is not necessary as a test of validity or objectivity of an observation or an evaluation. Rubbing bare observations together to produce sparks of correlations is a forlorn enterprise in much social inquiry. Rather, the basic reasoning pattern is closer to one of pattern-matching, of finding reasonable interpretations and explanations and understandings *within a given context.* The test of an explanation is not accuracy in predicting an

event but whether the audience can see new relations and answer "new but relevant" questions.

Finally, about the question of objectivity one must conclude one of two things: either objectivity cannot be exclusively identified with an externalized procedure totally separated from the minds that produced the observations and comprehended them; or else a great deal of truth is subjective in character. In the first case, objectivity means something more than it is commonly taken to mean; in the second case, it means something less.

What about validity? One definition of validity is that it is based on objective procedures. Validity carries with it the notions of being properly related to intent, of being correctly derived, and of being sanctioned by authority. In the narrow sense of quantitative objectivity, validity is equated with prediction—with checking the data against a criterion. But that assumes a single intent and assumes intersubjectivism as the verification principle. This is too narrow a procedure. Ultimately, says Cronbach (1971), validity is dependent on how the data are to be used and "utility depends upon values, not upon the statistical connections of scores."

If one cannot arrive at a single score presumably indicating validity, how is validity determined? Perhaps the best answer to the question is to examine the sources of invalidity. An evaluation may be invalid in a number of ways. One way is for the "facts and truths" upon which the evaluation is based to be wrong. Facts and truths are subject to the agreement of the universal audience. Many facts and truths are accepted without question by everyone. Other data must be determined by recognized data collection procedures, which are in turn sanctioned by a particular discipline and subject to public scrutiny. Often validity refers to using the accepted data collection procedures themselves, as Cronbach's article on test validation suggests.

Another way in which validity is at issue is in relating conclusions and interpretations to the data. As Cronbach asserts, it is not the test or the data collection procedures themselves so much as the interpretations that are valid or invalid. This is the validity of an inference. Is the inference correctly derived from the data and premises?

There is also the question of whether the interpretation can be properly applied to situations other than the one from which it was derived, since all generalizations are context dependent. These concerns have been dealt with in experimental design somewhat systematically as threats to internal and external validity.

In qualitative studies it is more difficult to provide evidence of validity—which is not a sign that it does not exist. Demonstrating validity in naturalistic studies usually consists of confirming one kind of data with another kind. In proposing case studies of science education, Stake and Easley

(1976) saw personal biases and past experience as the main threat to the credibility of the case studies. They proposed extensive tape recording of interviews, extensive use of direct quotations where possible, and reporting disagreements among respondents where they existed. People familiar with the local situation could read the written case to judge the accuracy of portrayal. The field workers would be keyed to "hints of inconsistency" for further pursuit. In instructions to on-site observers doing the studies, Stake (1976) urged confirming the observations through replication. Contexts for observations would be documented and elucidated. Securing the observations of several participants about a particular issue or event was a way of "triangulating" what actually happened.

Most of these threats to validity are seen from the perspective of a universal audience. But there is another way of looking at validity in evaluation—whether the evaluation is valid for particular audiences. After all, validity is always concerned with purpose and utility for someone. If the evaluation is not based on values to which the major audiences subscribe, these audiences may not see it as being "valid," i.e., relevant to them in the sense of being well-grounded, justifiable, or applicable. The evaluation may simply miss the main issues as far as particular audiences are concerned. At the same time the evaluation may be valid in the sense that the facts are correct and the inferences from the data correctly derived. From a particular audience's perspective, the premises may be the wrong ones.

An evaluation can also be invalid in this secondary sense if the argument forms employed are wrong. For example, in this society "means-ends" arguments, particularly cost-effectiveness arguments, are particularly potent. If one were to employ an argument based on maximizing excellence instead of choosing the best available alternative, it might carry little weight although being equally true and valid from the perspective of the universal audience. So validity can apply to evaluation in two rather different ways.

It is also the case that the more "naturalistic" the evaluation, the more it relies upon its audiences to draw its own generalizations (external validity). For example, a case study may be interpreted in different ways by each reader, since each reader has his own universe of cases in his mind for comparison. The reader can see similarities and differences based on his own experience and can draw his own interpretations.

Conceiving the process of generalization in this way alters even the first sense in which validity is used. The evaluator is still responsible for ascertaining and reporting "true" facts and statements, but part of the interpretation is beyond him. Since, as Cronbach says, the ultimate issue is the validity of the interpretation, which only the reader knows for sure, the audiences must assume considerable responsibility for the validity of their own interpretations. The evaluator must ultimately assume rational processes in the thinking of the audiences.

As Ennis (1973) noted, internal validity and external validity refer to rather different phenomena. External validity is concerned with the generalizability of general causal statements. Internal validity bears on specific causal statements that do not entail generalizing to new cases. Generalizing always assumes that one knows the relevant laws involved in extrapolating into new realms. An internally valid study, by contrast, only claims causality in the past within the specific circumstances. It claims no extrapolation and is hence less dependent on outside assumptions.

However, neither specific causal statements nor general causal statements follow perfectly logically from observations, even in the best experimental designs. Some empirical assumptions are needed even in the tightest design. In addition, identifying a particular event as a cause inescapably involves a judgment of responsibility that a particular event and no others is responsible for the effect, according to Ennis. This ascription of responsibility requires much background knowledge and a value judgment. It involves a probable assignment of praise or blame and suggests a place for intervention.

Most evaluators would assume responsibility for specific causal statements that "x caused y" in this study (internal validity), although this in itself necessarily involves a set of assumptions. But some would refer the generalizability of the findings to the audiences' judgments, since generalizability is based on outside information which the audiences but not the evaluator may have. The audiences might make some of the responsibility ascriptions based on their own background knowledge and values. Some evaluators, particularly naturalistic ones, might argue that this would ultimately result in superior generalizations.

There is yet a further related problem with objectivity. Is it really sufficient to say that an evaluator is objective? If objectivity is taken in the commonly-used sense of employing an externalized, specifiable procedure which produces replicable results, then it is certainly an insufficient criterion for an evaluation. The administration of standardized achievement tests is a totally externalized, specifiable procedure which produces replicable results. At the same time such tests are thought to be highly biased in many ways, particularly towards minority groups. In this sense, one has an objective but biased instrument. In fact one can produce an instrument in which the bias is in the other direction. (To further confound matters, if racial discrimination is the intent of such an instrument, one could have an objective, valid instrument for that purpose.)

An evaluation must be free from distortion and bias (qualitatively objective) and being externalized, specifiable, and replicable does not sufficiently address possible biases. Even qualitative objectivity is insufficient for evaluation, for it carries the aura of neutrality. People being evaluated do not want a neutral evaluator, one who is unconcerned about the issues. A person on trial would not choose a judge totally removed from his own social system.

Being disinterested does not give one the right to participate in a decision that determines someone's fate to a considerable degree. Knowledge of techniques for arriving at objective findings is inadequate. Rather, the evaluator must be seen as a member of or bound to the group being judged, just as a defendent is judged by his peers. The evaluator must be seen as caring, as interested, as responsive to the relevant arguments. He must be impartial rather than simply objective.

The impartiality of the evaluator must be seen as that of an actor in events, one who is responsive to the appropriate arguments but in whom the contending forces are balanced rather than non-existent. The evaluator must be seen as not having previously decided in favor of one position or the other.

The evaluator may resort to objective criteria to resolve the issues; but when his own impartiality is at stake, it is not enough that he give evidence of objectivity. He must give evidence of his impartiality by showing how he has acted contrary to his own interests in the past.

## EVALUATIVE DISCOURSE: THE GOOD LIFE (ALONG THE SAN ANDREAS FAULT)

It has been several weeks since I began this paper. The great Los Angeles earthquake has not yet come. Beautiful day succeeds beautiful day, each one much like the last; so it seems tomorrow must be like today, a pleasant dream extending indefinitely (argument by unlimited development).

Each day that passes makes the quake seem less likely than before. Yet if it is to occur this year, it should become *more* likely. I reason that the time I have remaining here is only a small fraction of the coming year, so the chances of the quake coming now are less than for the entire year of the prediction (argument by probability). I reason that even if the quake should come, the effects will not be disastrous (argument by consequences). In addition, the Midwest is racked by tornadoes (argument by comparison). Besides, would many of the smartest men in the country, including the seismologists, live here if the danger were so great (argument by incompatibility)? I feel reassured. My anxiety lessens.

Meanwhile within the last few days, the *New York Times Magazine* heightens the drama in its Bicentennial edition (July 4, 1976). As symbolic of "America at 200," it features a report on "The Good Life (along the San Andreas Fault)." On the cover is a painting of a fragment of a freeway jutting out into the empty ocean, the remains of Los Angeles after the next earthquake. The aricle begins with a six-paragraph scenario of the effects of the anticipated quake (arguments by symbol and illustration).

Those who literally live on top of the nine-mile deep fault have their own reasons for living there. As his backyard crumbles away daily, a postal worker, who has three cars, would like to move but cannot sell his

house (argument by sacrifice). A ranch manager who finds life better in California than anyplace he has ever lived explains, "I'm not leaving. Is there any place that doesn't have some catastrophe (argument by comparison)?"

For some the precariousness itself makes being here all the more precious. A dropped-out investment counselor living on the fault says, "You're living on a crisis point. Everything you have can be taken away from you at any time." More than any place, in every way, California is a challenge to the argument of unlimited development.

These are not the reasons I would give but they may be right. Each man is free to discover his own reasons. Each man is free to make his own choices. So it must be when faced with such uncertainty of knowing. Judgments cannot be based on an irrefutable reality. There will be a day when earthquakes are much more predictable than now. Even then, there will remain room for choice in how to respond. In social decision making certainty seems remote if not impossible.

Faced with such difficulty in arriving at an irrefutable reality, there are those who try to force simplicity atop the complexities of life and thereby eradicate ambiguity. They insist on pretending there is agreement where there is none, whether of facts or of values. Often in positions of power, they impose arbitrary definitions of reality for the sake of action. Yet reality is still there. Whatever even twenty-one million Californians believe, the great earthquake will come eventually.

The alternative is not necessarily a descent into irrationality. If opinions cannot be indisputably based, neither must they be regarded as entirely arbitrary, as being merely "value judgments." Such a classification identifies as knowledge only that which is clear, distinct, and unambiguous. This distinction establishes a schism between objectively true theoretical knowledge on the one hand and action based on irrational motives on the other. It culminates in designating as irrational those who do not agree with one's perspective. Classifying people as irrational justifies ignoring their opinions and perhaps their dignity and interests. It even legitimates using suggestion and force on them.

The alternative is to treat all men as rational. Between the conservative authoritarianism of tradition and the liberal authoritarianism of scientism, between the certainty of fanaticism and the irresponsibility of skepticism lies rational deliberation. One must take seriously the opinions of other people and engage them in serious discourse. This is the realm of argumentation and the proper sphere of evaluation.

The starting point is that groups of people adhere to opinions with variable intensity and that these beliefs can be put to the test of serious discourse. Even facts and values may be so considered. Rational discourse consists of giving reasons, although not compelling reasons. In the realm of action, where few things are clear and distinct, motivation can be rational. Practice can be reasonable.

The evaluator must engage his audiences in a dialogue in which they are free to employ their reasoning. This means that the audiences must assume personal responsibility for their interpretation of the evaluation since the reasoning presented to them is neither completely convincing nor entirely arbitrary. This means that the evaluator must also assume personal responsibility for his judgments since he cannot hide behind blind method. Both must exercise their natural reason.

## REFERENCES


Campbell, D. T. Assessing the impact of planned social change. In G. M. Lyons (Ed.), *Social research and public policies*. Dartmouth College, New Hampshire: The Public Affairs Center.

Campbell, D. T. Degrees of freedom and the case study. *Comparative Political Studies*, 1975, *8*(2).

Campbell, D. T., & Stanley, J. *Experimental and quasi-experimental design for research*. Chicago: Rand McNally, 1966.

Campbell, D. T. Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of egon brunswick*. New York: Holt, Rinehart & Winston, 1966.

Campbell, D. T. Qualitative knowing in action research. Presented at the American Psychological Association, New Orleans, 1974.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement*. Washington, D. C.: American Council on Education, 1971.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. Presented at American Psychological Association, New Orleans, 1974.

Denzin, N. K. The logic of naturalistic inquiry. *Social Forces*, 1971, *50*(2).

Ennis, R. On causality. *Educational Researcher*, 1973, *2*(6).

Gardner, M. Mathematical games. Scientific American, 1976, *234*(3).

Glass, G. V. Educational product evaluation. *Educational Researcher*, 1972, *1*(1).

Hamilton, D. A science of the singular? Mimeo. University of Illinois, 1976.

House, E. R. Justice in evaluation. In G. V. Glass (Ed.), *Evaluation studies—review annual*, Vol. 1, Beverly Hills, CA: Sage Publications, 1976.

Kemmis, S. Evaluation and evolution in knowledge about educational programs. Unpublished Doctoral Thesis. University of Illinois, 1976.

Kuhn, T. *The structure of scientific revolutions* (2nd ed.) Chicago: University of Chicago Press, 1970.

Levine, M. Scientific method and the adversary model: Some preliminary suggestions. *Evaluation Comment*, 1973, 4,(2), 1-3.

MacDonald, B. Evaluation and the control of education. University of East Anglia, May, 1974.

MacDonald, B., & Walker, R. Case-study and the social philosophy of educational research. University of East Anglia, August, 1974.

Madariaga, S. *Englishmen, Frenchmen, Spaniards*. London: Oxford University Press, 1949.

Mill, J. S. *A system of logic*. 1843. (8th ed.) New York: Harper, 1893.

Morgan, T. The good life (Along the San Andreas fault). *The New York Times Magazine*, July 4, 1976.

Ortony, A. Knowledge, language and teaching. Mimeo. University of Illinois, October, 1975.

Ortony, A. Why metaphors are necessary and not just nice. *Educational Theory*, Winter, 1975.

Owens, T. R. Educational evaluation by adversary proceedings. In E. R. House (Ed.), *School evaluation*. Berkeley: McCutchan, 1973.

Parlett, M., & Hamilton, D. Evaluation as illumination: A new approach to the study of innovatory programs. University of Edinburgh, October, 1972.

Perelman, C., & Olbrechts-Tyteca, L. *The new rhetoric: A treatise on argumentation.* University of Notre Dame Press, 1969.

Petrie, H. G. Metaphorical models of mastery: Or, How to learn to do the problems at the end of the chapter of the physics textbook. Mimeo. University of Illinois, 1976.

Polanyi, M. *Personal knowledge.* Chicago: University of Chicago Press, 1958.

Scriven, M. Educational product re-evaluation. *Educational Researcher,* 1972, *1*(5).

Scriven, M. Objectivity and subjectivity in educational research. *Philosophical Redirection of Educational Research.* National Society for the Study of Education, 1972.

Scriven, M. Goal-free evaluation. In E. R. House (Ed.), *School evaluation.* Berkeley: McCutcheon, 1973.

Scriven, M. Evaluation bias and its control. Occasional Paper 4. Kalamazoo: The Evaluation Center, Western Michigan University, June, 1975.

Shapley, D. Earthquakes: Los Angeles prediction suggests faults in federal policy. *Science,* May, 1976. *192.*

Smith, L. M., & Pohland, P. A. Education, technology, and the rural highlands. In R. E. Stake (Ed.), *Four evaluation examples: Anthropological, economic, narrative and portrayal.* AERA Monograph Series in Curriculum Evaluation, No. 7. Chicago: Rand McNally, 1974.

Stake, R. E. Program evaluations, particularly responsive evaluation. Goteborg, Sweden, 1973.

Stake, R. E. The case study method in social inquiry. University of East Anglia, February, 1976.

Stake, R. E., & Easley, J. Case studies in science education. A Proposal to the National Science Foundation, 1976.

Stake, R. E. Validation. Case Studies in Science Education Statement No. 24. Center for Instructional Research and Curriculum Evaluation, University of Illinois. Mimeo, October, 1976.

Strauch, R. E. A critical look at quantitative methodology. *Policy Analysis,* 1976, *2*(1).

Weizenbaum, J. *Computer power and human reason.* San Francisco: W. A. Freeman, 1976.

Wolf, R. L. The application of select legal concepts to educational evaluation. Unpublished doctoral dissertation, University of Illinois, 1974.

# APPENDIX

**Educational Product Evaluation: A Prototype Format Applied***

Gene V. Glass
Laboratory of Educational Research
University of Colorado

The conventions and techniques for evaluating educational products are not yet well established. Only recently have instructional materials and procedures been viewed as products to be developed and evaluated. Although the general procedures appropriate for evaluating consumer products are applicable to educational products, the unique characteristics of the education context raise special evaluation considerations. This paper addresses a "shelf-item" educational product that is of interest in its own right.

I. Product Description.

The product evaluated here is an instructional cassette recording "Evaluation Skills" (Tape 6B) created by Dr. Michael Scriven (Department of Philosophy, University of California, Berkeley) and produced by Dr. W. James Popham (School of Education, University of California, Los Angeles) for the American Educational Research Association (1126 16th St., N.W.; Washington, D.C.) under a grant from the U.S. Office of Education. The recording is intended primarily for in-service training of educational researchers and can be purchased for $6.00 from AERA.

The recording consists of a lecture on fundamental concepts of evaluation. The lecture is about 7,500 words long (the equivalent of approximately 17 single-spaced pages of typescript) and runs about 45 minutes. The 100-foot tape cassette can be played on any standard cassette player.

II. Goals Evaluation.

Product Goals are:

- To train educational researchers and others in the fundamentals of educational evaluation. The tape was commissioned ". . . to give the listener at least one important technical skill relating to educational research. . . . Although primarily intended as an update device for the educational researcher who has completed his formal training, many professors will find the tape ideal for their graduate classes." (*Educational Researcher*, Vol. 22, June 1971, p. 2)
- To provide an instructional product which can be used in situations (e.g., while driving) in which typical instructional products can't be used.
- To experiment with new instructional media.

There can be little quarrel with the first goal. Evaluation skills are in short supply. Legislation has created a significant demand for such skills, and a need for training in evaluation is commonly and justifiably expressed.

Making better use of otherwise dead time in commuting is commendable. The cassette tape is one of the few instructional media well suited to turning this unproductive time into something worthwhile. It is too soon to tell whether the ultimate, long-range effects of encroachment upon such private time will be undesirable. Nonetheless, it must be recognized that in extending an instructional

*Glass, Gene V. "Educational Product Evaluation: A Prototype Format Applied." *Educational Researcher*, January 1972, Vol. 1, No. 1, Pp. 7-10, 16. Copyright © 1972, American Educational Research Association, Washington, D.C.

Permission to reprint has been granted by AERA.

opportunity into time formerly not so used one may also be contributing to the destruction of peoples' senses of identity as persons apart from the roles they play as less than fully autonomous workers in huge, impersonal bureaucracies.

The goal of experimenting with new media is commendable to the degree that the choice of media for experimentation is made wisely (i.e., on the basis of data concerning costs, probable effectiveness, availability, longevity) and is not mere technological tinkering.

III. Clarification of Point of Entry of the Evaluator:

*Irreversible Decisions*

- USOE's grant to Popham (Program Director) to produce tapes.
- Popham's choice of topics and lecturers
- Lecturer's choice of subject-matter under the topic of "evaluation"
- AERA's reproduction of initial copies of the tape

*Reversible Decisions* (Enter the Evaluator)

- AERA's vending of initial copies
- AERA's choice of materials (cassette tapes)
- AERA's plans to sell additional copies of tapes in present form
- AERA's lack of plans to publish and distribute typescript

IV. Trade-Offs.

A series of trade-offs are involved in the production and application of this tape.

USOE traded off to produce the tape

- One-fourth of a 5-day training session for as many as 100 researchers
- The printing costs of 20,000 copies of 25 pages of prose materials for research training
- Half of one year's stipend for a doctoral-level educational research trainee
- Four all-expense scholarships for minority researchers to the AERA training session of their choice

The Cassette Tapes Project director traded off to produce the tape

- The production of typescript copies of the lectures
- The production of recorded synopses of several classic papers on educational evaluation

AERA continues to trade off to sell and produce the tape

- A small amount of managerial labor

The individual educational research would trade off to buy the tape

- Purchase of any four numbers in the AERA *Curriculum Evaluation Monograph Series.*
- Purchase of Wittrock-Wiley's *The Evaluation of Instruction*, or the April '70 issue of the *Review of Educational Research*, or *Suchman's Evaluative Research*, etc.
- Purchase of photo-copies of a half-dozen significant published papers on educational evaluation.

The trade-off with the greatest leverage that would retain the intent of the producer concerns the decision to produce and distribute the lecture as a cassette

**TABLE 1:**
**Cassette Recording Versus Typescript Costs**

| | |
|---|---|
| *Cassette Recording* | |
| 1. Production of master copy | |
| a. cost of tape only | $6.00 |
| b. cost of lecturer's services and expenses | $700. |
| 2. Reproduction of copies | |
| (*Cost of additional cassettes only*; no economy of scale) | $6.00/copy |
| 3. Mailing costs (4th class book rate) | $0.14 |
| 4. Operation of cassette recorder | |
| a. Purchase of recorder (price quoted on cheapest model) | $25.00 |
| b. Rental of recorder (rates range from $2.50 to $5.00 per day) | $3.75/day |
| c. operation of recorder | $0.00 |
| 5. Net cost of production and distribution of 100 cassette recordings (excluding lecturer's services) | $614.00 |
| *Typescript* | |
| 1. Production of master copy | |
| a. typing of 40 pages, double-spaced typescript | $8.00 |
| b. cost of lecturer's services and expenses | $700. |
| 2. Reproduction of copies (cost of paper and photocopying no economy of scale above 1,000 copies) | $0.40/copy |
| 3. Mailing costs (4th class book rate) | $0.14 |
| 5. Net cost of production and distribution of 100 typescripts (excluding lecturer's services) | $62.00 |

recording rather than a typescript. *Thus, the evaluation of the product will have a prominent comparative element in which a typescript of the lecture is the alternative product.*

V. Comparative Cost Analysis.

Table 1 summarizes the comparison, additional cost considerations follow:

*Simultaneous Mass listening.* For simultaneous teaching of 10–50 persons, the cassette recording could be economically used—even though there is significant distortion at higher volume on the Milovac (CR 203) Cassette Recorder.

*Tape costs.* The tape appears to be of high quality, perhaps too high since the voice frequencies of the tape do not require high fidelity reproduction. Since the lecture is only 45 minutes, it could have easily been recorded on a shorter, thicker, less expensive tape. There are other disadvantages of the thinner, more expensive tape: it tends to bind on cheaper players; print-through can occur in the recording process. It is presumed that nearly $5 was paid for each cassette. The evaluator has priced cassettes of acceptable quality at $0.75 per 60 minutes playing time (source: University of Colorado Bookstore). The entire cost of the tape cassette and reproduction from a master tape can be held below $2.00 (Authority: Dept. of Audio-Visual Instruction, Univ. of Colo.).

*Storage costs.* A 40-page typescript would occupy 65 $in^3$ of storage space compared to the 10 $in^3$ occupied by the cassette recording. If storage space became quite costly, the cost advantage would swing toward the cassette recording. However, under such circumstances the typescript could be transferred to *microfiche*,

for which storage (and usage) costs would be substantially below those of the cassette recording.

*Reducing costs for the typescript.* Prices for the typescript version of the lecture are quoted on a 40-page double-spaced manuscript. These costs could be significantly reduced by the following means: a) editing redundancies from lecture could reduce length by 10 per cent; b) single-spaced typing could reduce typescript length by almost half. Both a) and b) would result in a typescript version of the recording which could be sold for less than 20¢.

VI. Intrinsic (Secondary) Evaluation.

*Technical Quality*

Tape quality: Excellent. (But unnecessarily expensive.)

Recording fidelity: Excellent. The tape is free of background noise; volume is even.

Esthetic quality: Excellent. Lecturer's voice is well modulated; delivery is smooth and conversational.

Editing: Poor. Numerous stops-starts during recording (approximately a dozen) have garbled one or two words at the beginning of sentences; distracting and occasionally confusing. Approximately 10 seconds of recording is obscured at about the 80-foot mark of side 1.

Tape packaging: Poor. Sides (1 and 2) of tape are not marked. Cassette is difficult to remove from its poorly designed case. Erasure preventing devices on cassette were not activated by vendor. Label is not permanent and was poorly attached on the cassette purchased by the evaluator.

*Content Evaluation.*

1. Selection and Organization of Topics: Excellent
2. Use of Examples: Excellent
3. Clarity of Explanations: Excellent
4. Identification of Lecturer: Poor. Lecturer is identified only by name on label. No address or institutional affiliation is given for Lecturer even though he solicits communications from listeners at one point.
5. Accuracy of Scholarly Citations: Poor. Eisner volume is incorrectly cited as *Confronting Curriculum Evaluation.* Bloom, Hastings, Madaus handbook on formative evaluation is inadequately referenced as a "volume edited by Bloom." Wittrock & Wiley are cited, but authors' names are not spelled.

*Utilization of Uniqueness of Medium.*

The tape must be rated poor on this criterion. The lecturer claims that the opportunity to stop a recording is a unique feature of the medium ("the tape can be stopped more easily than the eye can be stopped from glancing ahead"); however, this claim cannot be substantiated in the opinion of the evaluator. Only about five requests for stops are made, and these requests are not very compelling. Furthermore they are probably inferior in eliciting thought when compared with adjunct questions in a typescript accompanied by answers at the end of the text.

The claim is also made that the tape can be played under circumstances in which reading is impossible or inconvenient (e.g., on airplanes or in cars). The range of circumstances in which the cassette recording is more convenient is probably smaller than the lecturer claims. Reading typescript on an airplane is quite conveniently done; furthermore, considerations of fellow travelers' comfort would

require an earphone, usually a no-cost but often-misplaced accessory. Whether the cassette recording will be utilized as is hoped (primarily in automobiles when no productive use of time would be made) remains to be seen. The data below bear on this possibility.

***Survey of Availability of Cassette Player and Incidence of Extended Commuting Among AERA Members.***

The following survey questionnaire was sent to a random sample of 100 members of AERA:

Dear AERA member

This survey is part of an evaluation of the AERA cassette tapes program. It is *not* sanctioned by AERA; they are not aware that it is being conducted.

We would appreciate your answering the following questions:

1. Do you have access to a cassette tape player (i.e., do you own one or could you borrow one at no cost)?

   Yes No

2. Do you commute by car to work for more than 20 minutes each way?

   Yes No

A total of 62 usable questionnaires were returned. The results permit the following conclusions regarding the availability of cassette players and their possible use while commuting to and from work:

1. Results. Frequencies of Response with Percents of Total Sampling.

| | | *Access to a cassette player* (yes) | (no) | |
|---|---|---|---|---|
| Commute more than 20 minutes each way to work | (yes) | 13 (21%) | 2 ( 3%) | 15 (24%) |
| | (no) | 39 (63%) | 8 (13%) | 47 (76%) |
| Totals | | 52 (84%) | 10 (16%) | 62 |

2. Conclusions.
   - That 84%[1] of AERA members have access to a cassette tape player indicates that AERA made a good choice of an "alternative" instructional medium.
   - Even though a substantial minority (20%)[1] of the AERA membership spends sufficient time commuting by car to make the tape medium of instruction advantageous, in terms of a head-count a substantial number (about 2000) of AERA members do commute under conditions which would permit instruction by cassette tapes.

VII. Outcome (Primary) Evaluation.

*Learning Rate.* Even if the aural medium is as effective for transmitting information as the visual medium (a question addressed later), it is undoubtedly slower. The speech rate for the cassette recording in question is approximately 160 words/minute (slightly slower than normal, conversational English). This is less than half what the reading rate would probably be for the typical listener (the average college freshman reads newspaper prose at more than 300 words/minute).

[1]These sample estimates are subject to substantial sampling error because of the small sample size (n = 62). The 95% confidence intervals on .84 and .24 are (.68, .92) and (.13, .40), respectively.

The effect on learning of this slower rate could be more serious than merely doubling the time required to learn the content of the recording. The slower rate of information presentation in the aural modality may tax the retentive powers of short-term memory to the extent that comprehension is seriously impaired.

A compressed speech version of the recording might correct problems allegedly associated with this low information transmission rate. Speech rates can be more than doubled by means of speech compressors without impairing comprehension. However, recording equipment may be prohibitively expensive.

*Provisions for Arbitrary Access.* Perhaps the principal disadvantage of recordings as a teaching device is that access to material on a tape at arbitrary points is awkward. Access to a particular section of a recorded lecture could be slower by a factor of ten or more than access to the same section in a typescript.

*Knowledge Acquisition in the Aural vs. the Visual Mode.* The relative efficiency of learning through visual and aural modalities has been debated in the history of psychology at least since 1894. As with most comparative educational research, the findings have been largely inconsistent and non-generalizable. Relative efficiency appears to depend on such interactive factors as 1) meaningfulness of the instructional material, 2) age of learner, 3) reading speed of learner, 4) intelligence of learner, 5) difficulty of the instructional material, and 6) whether retention is measured immediately or delayed. (For an excellent review of published studies on this question, see Travers, R.M.W. *et al. Research and Theory Related to Audio-Visual Information Transmission.* USOE Contract No. 3-20-003, 1967).

A recent experiment relevant to the comparative effectiveness of the cassette recording and typescript learning was performed by James R. Sanders (*Short-term and Long-term Retention Effects of Adjunct Questions in Aural Discourse*, Ph.D. thesis, Lab. of Ed. Research, Univ. of Colo., 1970). Sanders presented a 2000-word biography of William James to 72 undergraduates in either the visual or aural mode. Learning was measured immediately after presentation and one week later with a multiple-choice test. Results showed significantly ($p < .05$) greater learning in the visual mode (Sanders, 1970, p. 70).

VIII. Summative Judgments and Recommendations.

*Judgments*

The technical quality of the recording is good. The substantive content of the lecture is excellent. The recording is substantively more expensive than a typescript version of the same lecture and is probably less effective as a teaching device.

*Recommendations*

*To the individual researcher seeking to upgrade his understanding of evaluation:* Do not purchase this recording. Instead, buy *AERA Curriculum Evaluation Monograph* No. 1 and Suchman's *Evaluative Research* or purchase photocopies of the following papers:

Cronbach, L. J. "Evaluation for course improvement," *Teachers College Record*, 1963.

Scriven, M. "The methodology of evaluation," *AERA Curriculum Evaluation Monograph*, No. 1, Chicago: Rand-McNally, 1967.

Stake, R. E. "The countenance of educational evaluation," *Teachers College Record*, 1967.

If AERA offers for sale a typescript version of recording 6B (see Recommendations below), purchase it at any price up to $1.00 but not in place of purchasing

photo-copies of any of the above three papers or *AERA Curriculum Evaluation Monograph* No. 1 (Rand-McNally, 1967).

*To USOE:*

Cease allocating funds to the production of instructional recordings unless a compelling argument is presented that the instruction cannot be conducted in the visual mode (e.g., some instruction in music; training in auditory discrimination for young children; some instruction in speech pathology, linguistics, foreign language; "talking books" for the blind.)

Funds for training expended on development of products like that evaluated here would be better spent in support of the AERA Research Training Sessions program or in commissioning, reproducing, and disseminating training materials in typescript form.

*To AERA:*

Offer for sale at 75¢ per copy (to include mailing and handling) a typescript of the contents of Recording 6B. Offers for sale of the recording and the typescript should *not* be made separately.

Produce the cassette on cheaper tapes for the purchase whose circumstances make it an effective, superior learning device.

IX. Circumstances Modifying the Summative Judgments (Scope of the Value Claims).

The conclusion that the cost/effectiveness of the typescript version of the lecture is greater than the cost/effectiveness of the cassette recording would not be expected to hold (the superiority would be reversed) for sightless learners (who are also not deaf).

The cassette recording may be effective and is probably less expensive than the distribution of the typescript version for large groups (e.g., an undergraduate class) for which simultaneous mass listening is possible.

The cassette recording may be the only way to reach a segment of the population who might be characterized as "Reverse-Luddites" or "Mechanical Cultists," i.e., those persons who purchase electric carving knives, can openers, trail bikes, complex stereo systems, etc., and who claim—with vague appeals to McLuhan—that since books are passé they are no longer read.

X. Evaluating the Evaluator.

*Why an Evaluation?*

Gratuitous evaluation of products for which the net social investments are small can be a hostile act. Such "evaluations" can incur greater ultimate social costs than they reduce by destroying a sense of community among producers and evaluators, by creating defensiveness among producers who then refuse to cooperate with evaluators, by eroding civility in human relations, etc. *But in this case, the product developer asked to have his product evaluated.*

*The Evaluator's Motives.*

Evaluator's motives which would be served by a *favorable* over-all judgment: a) He is a member of the AERA Executive Board and would take satisfaction in the success of any AERA-sponsored activity. b) Persons involved in the production of the recording are colleagues of his and in a position indirectly to promote his general welfare.

Evaluator's motives which would be served by an *unfavorable* overall judgment: a) He declined an invitation to participate in the recording program on the grounds that it did not make use of the unique features of the medium and would not be cost/effective compared to dissemination of the presentations in written form. Hence, an unfavorable judgment would confirm his prejudgment and protect him against feeling that an opportunity had been lost. b) He was once beaten in a table-tennis match by the project director.

The evaluator has collected no representative data—either objective or subjective—on attitudes toward the product or its effectiveness as a learning device. His claim for the superiority of the typescript version of the lecture as a teaching device is based on extrapolation of the findings of a half-dozen experiments in audio-visual research comparing learning in the aural and visual modes.

## Educational Product Re-Evaluation*

Michael Scriven
University of California
Berkeley

1. Background

*(a).* The editor of *ER* invited this response to Glass' (Jan. '72) evaluation of the cassette I did for the AERA series concurrently with the acceptance of the Glass manuscript for publication.

*(b).* As Glass notes, I explicitly invited evaluations of the cassette and in fact offered a princely prize for the winning entry, namely, $8.00—the cost of the cassette. Glass' entry currently holds the lead in the competition for this prize, for the unimpeachably objective reason that it is the only one.

*(c).* It is of some interest that the production procedures of these cassettes involved one step of formative evaluation. Popham brought the authors to Los Angeles, where they uttered their talk into a microphone in a recording studio, without audience. The talk was recorded and also piped into a nearby room, where a number of experts and students heard it, and later critiqued it in a discussion with the author. In the light of this critique, the author then rewrought the talk at his leisure and taped it on a small portable which he was lent. The isolation of the recording act at Los Angeles was intended to simulate this final production procedure and to provide a chance to pick up technical deficiencies in recording procedure.

*(d).* While the formative evaluation of my performance was quite favorable, I decided to redo it completely. This involved some risks, not all of which paid off. It is, for example, possible that the new attempt was worse than the original one, and somewhat more probable that it was worse than a touched-up version of the original would have been. A second cycle of feedback would have been ideal, but was impractical. Three procedures are possible to handle such situations; *(i)* minimax strategy would support prohibition of "new starts"; *(ii)* funds might be budgeted to provide a second cycle in, say, 5% or 10% of the cases (I think I was the only such case); *(iii)* the producer might—as he did—take the chance that the author can improve his rating by making a fresh start. It would be interesting to have some data on these strategies.

2. Self-Evaluations

I have criticized the authors of the Phi Delta Kappa report on evaluation for a section in which they attempted a quasi-formal evaluation of their own book, *in* the book. I argued that if they came up with anything negative, they should revise the book; and if they didn't the endorsement was superfluous. That argument is oversimplified but still seems plausible. Now *reviewing* one's own book, as I once did by invitation is only one stage better; but there *is* a time lag and a chance for new critical input from others. *Replying* to reviews, as here, is then two stages from self-endorsement. The probability of bias has scarcely evaporated, but its probable direction is so obvious that it can do less harm than when concealed and there *is* a chance of useful rebuttals. The most interesting problem for the author in this role

*Scriven Michael. "Educational Product Re-Evaluation." *Educational Researcher*, May 1972, Vol. 1, No. 5. Inside front cover, pp. 2, 12. Copyright © 1972, American Educational Research Association, Washington, D.C.

Permission to reprint has been granted by AERA.

is distinguishing between giving excuses and rebutting criticisms. For example, one might excuse technical defects in the tape as due to equipment deficiencies, but this hardly affects the evaluator's complaint about them. Explanations may have some value for future projects of the same kind, but in attempting to achieve the best possible summative evaluation of the product, they are irrelevant. Since one's products are seen by others, as well as oneself, as extensions of oneself, it's very hard to avoid thinking of excuses as relevant. But I shall try to eschew them, and focus on summative product evaluation, as does Glass. Defense is therefore reserved against the usual inferences from the several deficiencies of product to those of the author or producer.

3. Specific Reactions

*(a). Technical—Hardware.*

Many of the comments made here seem correct, but one or two caveats should be considered. The value of good quality tape and cassettes is not immediately detectable. Both print-through and deterioration of signal-to-noise ratio under heating and magnetizing cycles increase with the years, and the mechanical components of a cassette are extremely susceptible to wear. The under $2.00 figure quoted by AVI at U. of Colorado indicates severely substandard materials. (Nevertheless, some saving might have been made here and it certainly would have been preferable if more of the tape could have been recorded, assuming any merit in the marginal material thereby added.) To the extent that teaching use is made of the tape, exposure to worn and over-magnetized heads and defective tape transports—common faults in classroom players—would increase the desirability of better quality materials. The distortion noted by Glass was due to the amplifier and speaker limitations of his player. Using an Advent 201 playing through MacIntosh electronics and AR transducers the results would, I judged, be quite satisfactory in a 2500-seat auditorium—even with an audience present. Of present portable players, Craig and Sony are pretty good products in the economy sector.

*(b). Technical—Software*

(Again, complaints not contested are conceded.) It seems likely that the deficiencies in citations would not significantly handicap usual library search routines. Feedback to me has proved possible for those who wrote c/o AERA or c/o Popham, not an excessively taxing procedure.

*(c). Crucial Comparisons*

The general procedure of really working to get estimates of comparative cost-effectiveness seems to me absolutely correct and indeed the method of choice in all educational evaluation. But the interpretation of the results by Glass may not be unimpeachable. Let me try shaking the kaleidoscope of data up a little to see how much stability there is in the image he reports. Consider the cassettes as serving solely these purposes: *(i)* Providing an improvement over listening to the car-radio, or music tapes, for drivers interested in educational research. (The use by those passengers in cars and on planes who find that reading in such circumstances gives them a headache is another 'exclusive' but small market use, which Glass does not identify.) *(ii)* Providing a cheap surrogate for a visiting lecturer in (approximately) graduate courses.

Now, it simply isn't interesting to compare cassettes with written materials vis-a-vis use *(i)*. No doubt we could all learn more than we do at home and in the office, but the AERA hasn't discovered a motivator yet, and until someone does,

it seems a useful service to offer an optional educational filling for the interstices of our life space. On cost and content there can be some serious criticisms, but the medium really has no competition when used as described. (Note that the cost is the same as that for commercial management training cassettes).

Use *(ii)* is usually a luxury use, of course. Written material has cost, speed, and replay advantages over tape. But it does not bring a new person into the didactic teaching process in quite the same way. Even getting to hear educational research personnel has some value in itself for the graduate student, as witness the reasons given for attending convention programs. There is also a possibility that the impact of several speakers will be stronger, motivationally, than one instructor plus readings. Again, an instructor in a particular classroom situation may feel the immediate importance of kicking in a diversion, a change of pace, an external authority. Without arguing for the general superiority of tape teaching, one can argue for its utility as a repertoire-enlarging device for special situations until its contribution can be shown to be *always* negligible. Is this an adequate justification for the use of the funds involved? That's a point-of-entry problem; it may be that the funds and/or enthusiasm were not available for any other production. Even if they were, the experimental commitment of AERA should justify *trying* a number of innovations like this one. The previous "success" of this in the medical in-service training area, and the management area, makes it a *reasonable* experiment, not a wild one. Of course, "success" in other fields has been subjectively and economically determined, not by proven learning gains. But if real tests are to be done, the strategy of doing it with AERA tapes and membership has a good deal to recommend it over doing it on medical tapes, for example, and trying to guess on extrapolation to educational researchers. So my principal criticism of the Glass evaluation concerns the choice of the main crucial comparison. It should not have been the typescript, but just the better content-cheaper package cassette. Broadly, evaluation should take care not to saddle the product with too large a "target population," one of the fallacies of "value dilution."

*(d). Use of Medium*

How could the content have been improved? There are many good answers to that, and Glass picks up several. I am not persuaded by his case for a poor rating on "use of the medium" however. To some extent, we are just trading hunches on this; I think it's harder to stop your eye skimming ahead on written material than it is to stop a tape, he does not. He thinks that my requests to stop are too few and not very compelling, etc. But I am not persuaded here, mainly because he does not suggest what *would* be good utilization of the medium. From my perspective, the most important factor is comprehensibility at listening speed, which Glass grants me under another heading. The interrogation idea was the only distinctive one I could think up. I expect there are others, but I'm not convinced that a "poor" rating on this dimension is justified until I see them.

4. Wider Horizons

*(a).* I was so impressed by Glass' willingness to do field research in the course of his evaluation, that I felt my response should also be based on a firm empirical foundation. Extensive field trials on a naive graduate student population has strongly confirmed my own belief in the existence of other populations besides those identified by Glass, or discussed above, for which this cassette may be useful.

Glass affirms utility for:

*(i)* "sightless learners (who are not also deaf)";
*(ii)* "large groups (e.g. an undergraduate class) for which simultaneous mass listening is possible";
*(iii)* " 'Reverse-Luddites' or 'Mechanical Cultists', i.e., those persons who purchase electric carving knives . . . trail bikes . . . etc., and who claim—with vague appeals to McLuhan—that since books are passé they are no longer read."

My survey indicates that group *(iii)* is too narrowly conceived. There really are normal people who prefer listening to discussions on the radio, over reading the transcript. And the cassette is controllable in a way the radio is not—no need for sudden dashes during commercials, for example. In the individual's diurnal prime, about 11 a.m., 2 p.m., and 7:30 p.m., reading works pretty well. But at the cyclic low points of the day, (7:45 a.m., 4 p.m., and 11:45 p.m.), there is a switch in optimal modality; a characteristic pattern of lying back with the eyes closed emerges, at which times auditory input is quite acceptable. Further details of this study must await replication, which I shall attend with confidence that some of the minor deficiencies in what is, after all, a pilot study ($n = 1$) are more than compensated by the quality of the naive graduate student population (my spouse).

*(b).* Evaluation of educational products frequently, but understandably, overlooks the learning pay-off for the intermediary population, usually the teacher. In this case the producer (probably) and lecturer (certainly) have learnt a great deal from producing this cassette. This is a small group, but one with potentiality for significant further effect on the theory and practice of evaluation. A substantial part of this learning has come from the evaluation of the cassette by Glass.